A Hybrid Approach for Radio Access Technology Selection in Heterogeneous Wireless Networks

Melhem El Helou, Samer Lahoud, Marc Ibrahim, Kinda Khawam, Bernard Cousin, Dany Mezher

Noname manuscript No.
(will be inserted by the editor)

A Hybrid Approach for Radio Access Technology Selection in Heterogeneous Wireless Networks

Melhem El Helou · Samer Lahoud · Marc Ibrahim · Kinda Khawam · Bernard Cousin · Dany Mezher

Received: October 22, 2014 / Accepted: July 16, 2015

Abstract In heterogeneous wireless networks, different radio access technologies are integrated and may be jointly managed. To optimize network performance and capacity, efficient Common Radio Resource Management (CRRM) mechanisms need to be defined. This paper tackles the Radio Access Technology (RAT) selection, a key CRRM functionality, and proposes a hybrid decision framework that dynamically integrates operator objectives and user preferences. Mobile users are assisted in their decisions by the network that broadcasts cost and QoS information.

Our hybrid approach involves two inter-dependent decision-making processes. The first one, on the network side, consists in deriving appropriate network information so as to guide user decisions in a way to meet operator objectives. The second one, where individual users combine their needs and preferences with the signaled network information, consists in selecting the RAT to be associated with in a way to maximize user utility.

We first focus on the user side and present a satisfaction-based multi-criteria decision-making method. By avoiding inadequate decisions, our algorithm outperforms existing solutions and maximizes user utility. Further, we introduce two heuristic methods, namely the staircase and the slope tuning policies, to dynamically derive network information in a way to enhance resource utilization. The performance of each decision-making process, on the

Melhem El Helou, Marc Ibrahim and Dany Mezher
Ecole Supérieure d'Ingénieurs de Beyrouth (ESIB), Faculty of Engineering, Saint Joseph University of Beirut, Beirut 1107 2050, Lebanon E-mail: melhem.helou@usj.edu.lb; marc.ibrahim@usj.edu.lb; dany.mezher@usj.edu.lb

Samer Lahoud and Bernard Cousin
Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA) Laboratory, University of Rennes 1, Rennes 35042, France E-mail: samer.lahoud@irisa.fr; bernard.cousin@irisa.fr

Kinda Khawam
PRiSM Laboratory, University of Versailles, Versailles 78035, France E-mail: kinda.khawam@prism.uvsq.fr

network and user sides, is evaluated separately through extensive simulations. A comparison of our hybrid approach with six different RAT selection schemes is also presented.

Keywords Radio access technology selection · heterogeneous wireless networks · hybrid decision-making · QoS · resource utilization

1 Introduction

Along with the rapid growth of mobile broadband traffic, different radio access technologies including 3GPP families such as Universal Mobile Telecommunications System (UMTS), High Speed Packet Access (HSPA), and Long-Term Evolution (LTE) and IEEE families such as WiFi and Worldwide Interoperability for Microwave Access (WiMAX) are being integrated. When their radio resources are jointly managed, they form a heterogeneous wireless network able to provide high capacity and cost-effective global service coverage (Fig. 1).

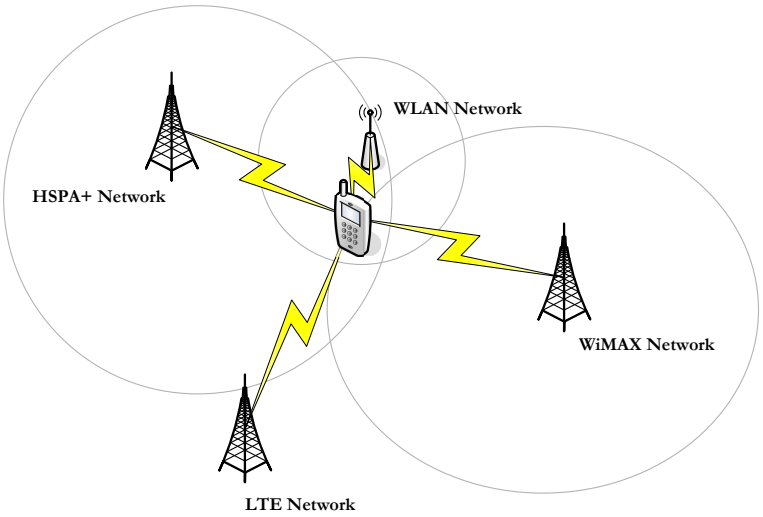


Fig. 1 A typical heterogeneous wireless network

To optimize network performance while enhancing user experience (Always Best Connected concept [18]), efficient Common Radio Resource Management (CRRM) mechanisms [31] need to be defined. Typically, when a new or a handover session arrives, a decision must be made as to what technology it should be associated with. This is known as the Radio Access Technology (RAT) selection, a key CRRM functionality.

So as to consider operator objectives, including efficient resource utilization, network-centric schemes have been proposed: network elements collect necessary measurements and information. They take selection decisions transparently to end-users in a way to enhance heterogeneous network performance.

However, to reduce network complexity, signaling and processing load, mobile-terminal-centric methods have also gained in importance. Based on their individual needs and preferences, rational users select their RAT in a way to selfishly maximize their utility. Because individual users have no information on the global network state (*i.e.,* network load conditions), mobile-terminal-centric approaches are known for their potential inefficiency. Although mobile users try to selfishly maximize their interest, their selection decisions are eventually in no one long-term interest. This dilemma is known as the *Tragedy of the commons* [19].

The challenge is then to design a RAT selection method that jointly enhances overall network performance and individual user experience, without unduly complicating the network. Thus, in this article, we propose a new hybrid decision method that combines benefits from both network-centric and mobile-terminal-centric approaches. The network information, that is periodically broadcasted, assists mobile users in their decisions. More precisely, individual mobiles select their RAT based on their needs and preferences as well as on the cost and partial QoS parameters signaled by the network. By broadcasting appropriate decisional information, the network tries to globally control user decisions in a way to meet operator objectives. This hybrid framework may be naturally integrated into Self-Organizing Networks (SON) [2]: network parameters are automatically tuned so as to self-optimize serving RATs.

When several base stations are available, decisions are traditionally based on received-signal-strength measurements. In our work, so as to maximize user experience, we present a satisfaction-based Multi-Criteria Decision-Making (MCDM) method. In addition to their radio conditions, mobile users consider cost and QoS parameters signaled by the network, when selecting their RAT. In comparison with existing MCDM solutions, our algorithm meets user needs (*e.g.,* traffic class, throughput demand, cost tolerance), avoiding inadequate decisions. A particular attention is then addressed to the network to make sure it broadcasts appropriate decisional information so as to better exploit its radio resources, while mobiles are maximizing their own utility. We thereby present two heuristic methods, namely the staircase and the slope tuning policies, to dynamically derive what to signal to mobiles. We also investigate the performance improvement achieved by providing differentiated service classes and minimum bandwidth guarantees to mobile users, regardless of future network load conditions. These implementation choices help to significantly enhance user experience and operator gain.

2 Related Work

Heterogeneous networks have triggered considerable interest among researchers in the past few years. Several papers have addressed the RAT selection problem. In [14], the selection decision is isomorphically mapped to a multiple choice multiple dimension knapsack problem, known to be NP-hard.

In [33,23,32], RAT assignment is formulated as an optimization problem. Exact and heuristic algorithms are used to derive an optimal or a near optimal solution, that optimizes the global network utility (*e.g.,* user-perceived throughputs, service times). In [26], RAT selection and resource allocation are simultaneously performed. The proposed CRRM algorithm considers the discrete nature of mobile radio resources and is then based on integer linear programming optimization techniques. Radio resources are distributed in a way to maximize network capacity, providing users with satisfactory QoS levels. In [16], based on the CEA (Constrained Equal Awards) bankruptcy rule, selection decisions try to equally satisfy mobile users: they are assigned the same amount of resources, without exceeding their individual demands. In [20,6,39,25,7,34,42,44,43], a Semi-Markov Decision Process (SMDP) is employed to model the RAT selection decision-making. A set of states, actions, rewards, and transition probabilities are defined. Linear or dynamic programming algorithms are adopted to find the optimal access policy that maximizes the long-term reward function (*i.e.,* the expected utility calculated over an infinitely long trajectory of the Markov chain). In [15], a fuzzy neural methodology is proposed to jointly decide of the RAT selection and the bandwidth allocation. A reinforcement signal is generated to optimize the decision-making process: the means and the standard deviations of the input and output bell-shaped membership functions are adjusted accordingly. As network elements gather information about individual users, namely their QoS needs, and their radio conditions in the different serving cells, network-centric approaches generally optimize resource utilization. Yet, network complexity, processing, and signaling load are drastically increased.

To face the high computational complexity of network-centric methods, mobile-terminal-centric heuristics are proposed in [27]. Distance-based, probabilistic distance-based, peak rate-based and probabilistic peak rate-based algorithms are presented: they indicate the probability to assign mobiles to the primary (IEEE 802.11g) and to the secondary (IEEE 802.11b) technologies based on their distance from the two access points or on the peak rate they can achieve when connected to these access points. Since user payoff does not only depend on its own decision, but also on the decisions of other mobiles, game theory is used as a theoretical framework to model user interactions in [3,23,21,30,25]. Players (*i.e.,* the individual users) try to reach a mutually agreeable solution, or equivalently a set of strategies they unlikely want to change. Yet, the convergence time seems to be long [23]. In [36,9,8], RAT selection is formulated as a reinforcement learning problem. A set of states, actions, and rewards are defined. Mobiles iteratively learn selection decisions, through trial-and-error interaction with their environment, in a way to maximize their utility. They discover a variety of actions, and progressively favor effective ones. In [29,5,13,38,4,35] multi-criteria decision-making methods, including Simple Additive Weighting (SAW), Multiplicative Exponent Weighting (MEW), Grey Relational Analysis (GRA) and Technique for Order Preference by Similarity to Ideal Solution (TOPSIS), are employed. Individual users combine their QoS parameters (*e.g.,* instantaneous peak rates), calcu-

late decision metrics, and select their RAT accordingly. In [41,13,5], fuzzy logic is also used to deal with the imprecise information of some criteria and user preferences. As mobiles autonomously select their RAT, network operations remain reduced. Furthermore, decisions can easily involve user needs and preferences, and various mobile-terminal-related parameters. However, when mobiles do not cooperate, mobile-terminal-centric approaches potentially lead to performance inefficiency.

In order to avoid the drawbacks of network-centric and mobile-terminal-centric approaches, we propose in this paper a hybrid decision solution that:

- minimizes network complexity, signaling and processing load: a common network information is periodically broadcasted using the logical communication channel (*i.e.,* radio enabler) proposed by the IEEE 1900.4 standard [1]. Selection decisions are, however, left to the mobiles.
- efficiently utilizes radio resources despite of the non-cooperative behavior of mobile users: by broadcasting appropriate decisional information, the network tries to guide user decisions in a way to satisfy operator objectives (*e.g.,* enhance resource utilization).

Our hybrid approach actually involves two decision-making processes. The first one, on the network side, consists in deriving appropriate network information so as to guide user decisions in a way to meet operator objectives. The second one, where individual users combine their needs and preferences with the signaled network information, consists in selecting the RAT to be associated with in a way to maximize user utility. Since, in their turn, individual user decisions influence the upcoming network information, the two decision-makings are considered to be inter-dependent. Thus, selection decisions dynamically integrate both operator objectives and user needs and preferences.

Each decision-making, on the network and user sides, is studied and evaluated separately. We also present a comparison of our hybrid approach with different network-centric, hybrid and mobile-terminal-centric solutions.

The rest of this paper is organized as follows: Section 3 describes our hybrid decision framework. Our satisfaction-based multi-criteria decision-making method is presented in section 4. Section 5 introduces our tuning policies. Section 6 provides a detailed performance evaluation of the two decision-making processes. In section 7, we compare our hybrid approach with six different RAT selection schemes, including network-centric, hybrid and mobile-terminal-centric methods. Section 8 concludes the article.

3 Hybrid Decision Framework

In this section, we first present our network model: network topology and radio ressources. Then, we describe our hybrid decision framework: what network information is sent to all mobiles, and how mobiles select their serving RAT.

3.1 Network topology

We consider a heterogeneous wireless network composed of N_T RATs. The modulation and coding scheme, that can be assigned to a user connected to RAT x, differs depending on its radio conditions in the cell, more precisely on its signal-to-noise ratio denoted by SNR^x. As the number of possible modulation and coding schemes is limited, we decompose the cell into N_Z^x zones with homogeneous radio characteristics [20,6,7]. Users in zone Z_k^x, $k = 1, ..., N_Z^x$, are assumed to have a signal-to-noise ratio between δ_k^x and δ_{k-1}^x, and then to use $mod^x(k)$ with $cod^x(k)$ as modulation and coding scheme:

$$(mod^x(k), cod^x(k)) = \begin{cases} \text{none} & \text{if } SNR^x(k) < \delta_{N_Z^x}^x, \\ (mod_{N_Z^x}^x, cod_{N_Z^x}^x) & \text{if } \delta_{N_Z^x}^x \leq SNR^x(k) < \delta_{N_Z^x-1}^x, \\ \ldots & \\ (mod_1^x, cod_1^x) & \text{if } \delta_1^x \leq SNR^x(k) < \delta_0^x = \infty. \end{cases} \quad (1)$$

where $\delta_{N_Z^x}^x$ is the minimum signal-to-noise ratio, that allows transmission at the lowest throughput, given a target error probability.

Furthermore, and for the sake of simplicity, users in a same zone are assumed to have the same peak throughput, realized when present alone in the cell.

In the remainder, let the N_Z^x-tuple $n^x = (n^x(k))$, for $k \in \{1, ..., N_Z^x\}$, be the state of RAT x. $n^x(k)$ represents the number of users, in zone Z_k^x, that are connected to RAT x. The state s of the heterogeneous wireless network is the concatenation of RAT x substates, for $x \in \{1, ..., N_T\}$: $s = (n^x)$, for $x \in \{1, ..., N_T\}$.

3.1.1 Cell Decomposition

Because of fading effects, radio conditions are time-varying. User signal-to-noise ratio can take all possible values, leading to different modulation and coding schemes. However, as RAT selections are made for a sufficiently long period of time (*e.g.,* session duration, user dwell time in the cell), users are distributed over logical zones depending on their average radio conditions, rather than on their instantaneous ones.

Another approach is found in [20], where an analytical radio model, that accounts for interference, path loss, and Rayleigh fading, is used. It has been demonstrated that users need to be situated at $r_k \in [R_{k-1}^x, R_k^x[$ from their base stations, so as to have a signal-to-noise ratio between δ_k^x and δ_{k-1}^x, with at least a high probability $\mathbb{P}_{th}$. This means that the cell may be divided into concentric rings, as illustrated in Fig. 2, and mobiles in ring Z_k^x will use $mod^x(k)$ with $cod^x(k)$ as modulation and coding scheme, with at least a high probability $\mathbb{P}_{th}$. Further, to define the different rings, the distances R_k^x have been analytically derived, mainly as a function of δ_k^x, $\mathbb{P}_{th}$, and radio model parameters.

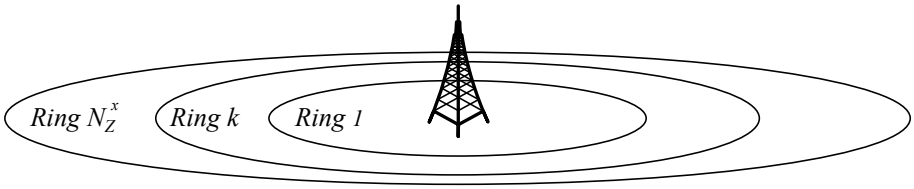


Fig. 2 RAT x cell divided into N_Z^x concentric rings

3.2 Network Resources

Prior to the RAT selection process, a common admission control is assumed to be performed. New and handover sessions are admitted to the extent that joint available resources are able to meet their requirements, while not compromising the QoS level of ongoing ones. Further, after sessions are accepted, decisions are made as to what RAT they should be associated with. Robust decisions are crucial to avoid network congestion, and enhance user experience.

In RAT x, the radio resource is divided into elementary resource units (RU). Typically, in OFDM(A)-based technologies (*e.g.,* LTE and WiMAX), resource units are defined as OFDM symbols (one-dimensional allocations), or OFDMA slots (two-dimensional allocations: m subcarriers by n OFDMA symbols). However, in CDMA-based technologies (*e.g.,* HSPA), codes, power and allocation times are regarded as RUs.

In the time domain, transmissions are organized into radio frames of length T^x. At each scheduling epoch, RUs are allocated to individual users, based on a predefined scheduling algorithm. User throughputs depend on their allocated RUs (*i.e.,* their description and amount), and modulation and coding schemes. Typically, when fair time scheduling is employed, cell resources (*e.g.,* codes, power and allocation times in HSPA, OFDMA slots in LTE) are equally distributed to mobile users [37]. Yet, mobiles with good radio conditions (*e.g.,* cell center users) experience a higher throughput than those with bad radio conditions (*e.g.,* cell edge users).

3.3 Network Information

Periodically or upon user request, network information is sent to all mobiles, using the logical communication channel (*i.e.,* radio enabler) proposed by the IEEE 1900.4 standard [1]. This logical channel allows information exchange between the Network Reconfiguration Manager (NRM) on the network side, and the Terminal Reconfiguration Manager (TRM) on the mobile-terminal side (Fig. 3). The purpose is to improve resource utilization and user experience in heterogeneous wireless networks.

In our work, by appropriately tuning network information, the network globally controls user decisions, in a way to meet operator objectives (*e.g.,* enhance network performance, minimize energy consumption). Network infor-

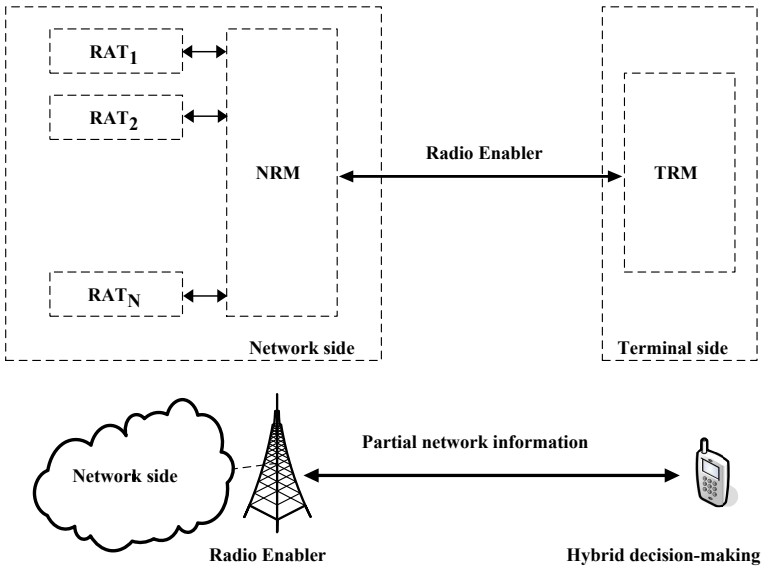


Fig. 3 Hybrid 1900.4 network architecture

mation may then be static or dynamic, so as to optimize short- or long-term network utility.

When a new or a handover session arrives, the mobile decodes network information, evaluates serving RATs, and selects the one that maximizes its own utility. As a matter of fact, selection decisions depend on user needs and preferences, as well as on the signaled network information.

The network is fully described by its state s. Yet, in our work, only monetary cost and partial QoS parameters are sent to mobiles. This reduces signaling load. Furthermore, by masking RAT load conditions, QoS information may reflect not only the current network state s, but also other network-related parameters (*e.g.,* energy consumption). For instance, QoS parameters may be tuned, so that mobile decisions are consistent with operator energy-saving objectives. This flexible design allows the network to derive cost and QoS parameters in a way to optimize a generic utility function.

Moreover, cost and QoS parameters, signaled by the network, are seen as incentives to join serving RATs:

- Cost parameters: Because flat-rate pricing strategies waste resources [10], result in network congestion and thus degrade network performance [40], they are not optimal in supporting QoS. A volume-based model is therefore proposed: mobile users are charged based on the amount of traffic they consume. In our work, *costs* are defined on a per kbyte basis.
- QoS parameters: The amount of resource units (RUs) to be allocated to future arrivals are broadcasted:
  - Mobiles are guaranteed an average minimum amount of RUs, denoted by n_{min}.

– They also have priority to occupy up to an average maximum amount of RUs, denoted by n_{max}.

Because the smallest allocation unit (*i.e.,* RU) has different descriptions in the different RATs, there is a need to homogenize the QoS information. QoS parameters are then expressed as throughputs: d_{min} and d_{max} instead of n_{min} and n_{max}. However, as user throughputs strongly depend on their radio conditions, d_{min} and d_{max} are derived for the most robust modulation and coding scheme (*i.e.,* $mod^x_{N^x_Z}$ with $cod^x_{N^x_Z}$).
Therefore, when evaluating serving RATs, mobiles should combine their individual radio conditions with the provided QoS parameters: for that, they multiply d_{min} and d_{max} with a given modulation and coding gain, denoted by $g(M, C)$.

Although QoS parameters are provided, our decision framework is independent of local resource allocation schemes. First, the minimum guaranteed RUs, namely n_{min}, are directly granted. Then, any priority scheduling algorithm, including opportunistic schemes [22,17,24,28], could be adopted to share out remaining resources. Grants are, however, limited to n_{max}. Residual resources are afterwards equitably distributed: when all mobiles have received their maximum throughput, they are considered to have the same priority, leading to fair allocation.

3.4 RAT Selection

The network proposes one or more alternatives, that are the available RATs. For each alternative a, the network broadcasts the three parameters: $d_{min}(a)$, $d_{max}(a)$, and $cost(a)$. From the user point of view, these parameters are the decision criteria to be used to evaluate serving RATs. As in all multi-criteria decision making methods, mobiles define and compute a utility function for all of the available alternatives. This utility is obtained after normalizing and weighting the decision criteria.

4 Satisfaction-based Decision Method

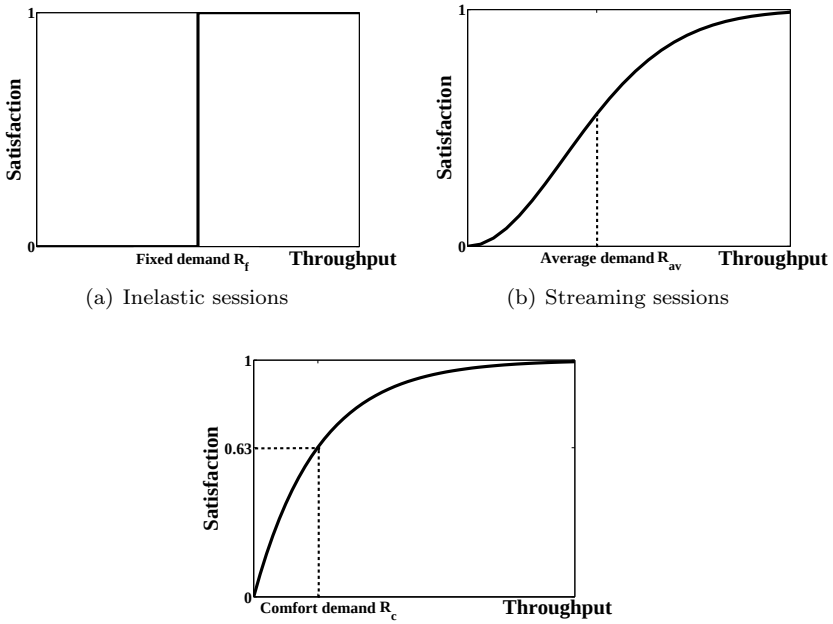
In this section, we present our Satisfaction-Based (SB) Multi-Criteria Decision-Making (MCDM) method. The particularity of our algorithm resides in the normalization step, that takes into account user needs (*i.e.,* traffic class, throughput demand, cost tolerance). By avoiding inadequate decisions, our algorithm overcomes some limitations of well-known MCDM methods.

4.1 Normalization and Traffic Classes

The normalization of the decision criteria $d_{min}(a)$, $d_{max}(a)$, and $cost(a)$ takes into consideration session traffic class, throughput demand, and cost toler-

ance. For traffic class c and alternative a, $\hat{d}^c_{min}(a)$, $\hat{d}^c_{max}(a)$, and $\widehat{cost}^c(a)$ are respectively the normalized values of $d_{min}(a)$, $d_{max}(a)$, and $cost(a)$.

In our work, we define three traffic classes : inelastic, streaming, and elastic classes. Before we give the normalizing functions for each traffic class, we note that $\hat{p}^c(a), p \in \{d_{min}, d_{max}, cost\}$, can be viewed as the expected satisfaction of a class c session, with respect to criterion p, when alternative a is selected:



(a) Inelastic sessions

(b) Streaming sessions

(c) Elastic sessions

Fig. 4 Throughput satisfaction function forms

– Inelastic sessions ($c = I$): since designed to support constant bit rate circuit emulation services, inelastic sessions require stringent and deterministic throughput guarantees. d_{max} should have no impact on RAT selections. Besides, the satisfaction with respect to d_{min} has a step shape (Fig. 4(a)). When alternative a is selected, mobiles expect to be satisfied provided that their minimum guaranteed throughput $d_{min} = d_{min}(a) \cdot g(M, C)$ is greater or equal to their fixed throughput demand R_f; otherwise, they are not satisfied.

$$\hat{d}^I_{min}(a) = \begin{cases} 0 & \text{if } d_{min}(a) \cdot g(M,C) < R_f \\ 1 & \text{if } d_{min}(a) \cdot g(M,C) \geq R_f \end{cases} \quad (2)$$

– Streaming sessions ($c = S$): since designed to support real-time variable bit rate services (*e.g.,* MPEG-4 video service), streaming sessions are fairly

flexible, and usually characterized by a minimum, an average and a maximum throughput requirement. Therefore, when alternative a is selected, their expected satisfaction with respect to d_{min} and d_{max} is represented by an S-shaped function (Fig. 4(b)):

$$\hat{d'}^{S}(a) = 1 - \exp(\frac{-\alpha(\frac{d'(a).g(M,C)}{R_{av}})^2}{\beta + (\frac{d'(a).g(M,C)}{R_{av}})}) \quad (3)$$

where $d' = \{d_{min}, d_{max}\}$.
R_{av} represents session needs: an average throughput demand. α and β are two positive constants necessary to determine the shape of the S-shaped function.

– Elastic sessions ($c = E$): since designed to support traditional data services (*e.g.,* file transfer, email and web traffic), elastic sessions typically using the TCP protocol adapt to resource availability. As they require no QoS guarantees, d_{min} has no impact on RAT selections. Moreover, the satisfaction with respect to d_{max} has a concave shape as illustrated in Fig. 4(c).
User satisfaction is expected to increase slowly as its throughput exceeds its comfort throughput demand R_c (*i.e.,* the mean throughput beyond which user satisfaction exceeds 63% of maximum satisfaction):

$$\hat{d}^{E}_{max}(a) = 1 - \exp(-\frac{d_{max}(a).g(M,C)}{R_c}) \quad (4)$$

Furthermore, the monetary cost satisfaction is modeled as a Z-shaped function for all sessions (Fig. 5): the slope of the satisfaction curve increases rapidly with the cost.

$$\widehat{cost}^{c}(a) = \exp(-\frac{cost(a)^2}{\lambda^c}), c \in \{I, S, E\} \quad (5)$$

λ^c represents the cost tolerance parameter: a positive constant to determine the shape of the Z-shaped function.

4.2 User Profile and Utility Function

The user profile defines the cost tolerance parameter and the weights to be applied to normalized criteria. More precisely, the user profile is the set of vectors $(\lambda^c, w^c_{d_{min}}, w^c_{d_{max}}, w^c_{cost}), c \in \{I, S, E\}$, where w^c_p is the weight of $\hat{p}^c, p \in \{d_{min}, d_{max}, cost\}$. When alternative a is selected, the expected utility of a class c session is defined as follows:

$$U^c(a) = w^c_{d_{min}} \cdot \hat{d}^c_{min}(a) + w^c_{d_{max}} \cdot \hat{d}^c_{max}(a) + w^c_{cost} \cdot \widehat{cost}^{c}(a)$$

Note that predefined user profiles (*e.g.,* cost minimizing profile, QoS maximizing profile) may be introduced. Thereby, end-users do not worry about technical details: they can use default values for the cost tolerance parameter, and the decision criteria weights.

Fig. 6 summarizes the decision process:

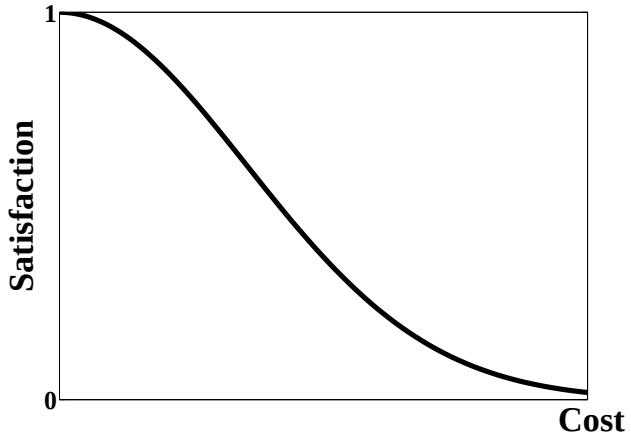


Fig. 5 Monetary cost satisfaction function, ($\lambda^c = 25$)

- For each alternative a, the mobile combines its radio conditions with the QoS parameters signaled by the network: it multiplies $d_{min}(a)$ and $d_{max}(a)$ with a given modulation and coding gain, to determine its perceived QoS parameters, as provided by the network.
- Then, based on user needs (*i.e.,* traffic class c, throughput demand and cost tolerance λ), it computes the normalized decision criteria: $\hat{d}^c_{min}(a)$, $\hat{d}^c_{max}(a)$ and $\widehat{cost}^c(a)$.
- Next, it combines user preferences (*i.e.,* $w^c_{d_{min}}$, $w^c_{d_{max}}$ and w^c_{cost}) with the normalized decision criteria, so as to compute the weighted normalized criteria: $w^c_{d_{min}} \cdot \hat{d}^c_{min}(a)$, $w^c_{d_{max}} \cdot \hat{d}^c_{max}(a)$ and $w^c_{cost} \cdot \widehat{cost}^c(a)$.
- Finally, it computes the utility function for each alternative a, and selects the one with the highest score.

This decision process is performed at session initiation and possibly also during session lifetime. Mobiles decide of their serving RAT based on their individual needs and preferences, as well as on the broadcasted network information. However, they can migrate to another RAT following changes in their radio conditions. At this point, mobiles check whether their serving RAT is still their best choice, or in other words, whether it is still expected to maximize user utility. An inter-RAT handover is triggered only when another RAT can provide users with significantly higher satisfaction level. This helps to reduce unnecessary handovers (*i.e.,* ping-pong effect).

5 Tuning Policies

Because mobile users also rely on their needs and preferences when selecting their RAT, the network does not completely control individual decisions. Yet,

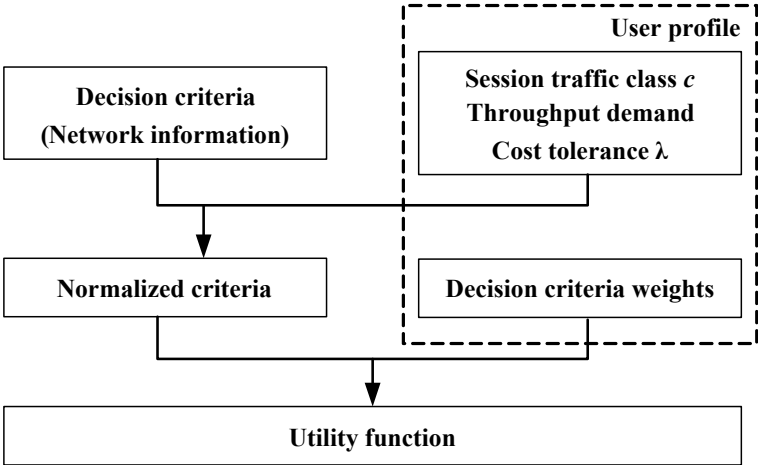


Fig. 6 Satisfaction-based multi-criteria decision process

by signaling appropriate decisional information, the network tries to globally guide user decisions, in a way to meet operator objectives. These may include energy savings: mobiles are pushed to some base stations, while others are switched to sleep mode so as to save energy. In our work, we assume that operators are only concerned by efficiently utilizing their radio resources: providing better network performance, higher user satisfaction, and larger operator gain.

When a RAT dominates all the others (*i.e.,* provides higher QoS parameters for the same cost, or the same QoS parameters for a lower cost), common radio resources are inefficiently utilized, causing performance degradation. In fact, mobile users would select the dominant alternative, leading to unevenly distributed traffic load. While a RAT is overcrowded, the others are almost unexploited. This inefficiency is very similar to that of the mobile-terminal-centric approaches. To avoid it, QoS parameters, signaled by the network, needs to be modulated as a function of the load conditions.

In this section, we present two heuristic methods, namely the staircase and the slope tuning policies, to dynamically derive QoS information. In order to reduce network complexity and processing load, one of the drawbacks of network-centric approaches, our policies are made simple. Yet, they help to efficiently distribute traffic load over the available RATs, and thus to better utilize radio resources.

5.1 Staircase Tuning Policy

The load factor represents the amount of throughput guarantees, and is defined as the ratio of the number of guaranteed allocated RUs to the total number of RUs. Fig. 7 illustrates how QoS parameters, namely d_{min} and d_{max} separately, are tuned as a function of the load factor using the Staircase policy. When RAT x load factor is low, the network can promise high throughput guarantees to

arriving mobiles to join RAT x. The highest $d_{min}(x)$ and $d_{max}(x)$ values are signaled. However, as RAT x load factor exceeds S_1 threshold, the network needs to reduce QoS incentives in RAT x so as to avoid RAT x congestion, or in other words, to avoid resource shortage in RAT x. QoS parameters are separately decreased, following a step function. Moreover, as S_2 is reached, the network no longer provides incentives to arriving mobiles in RAT x.

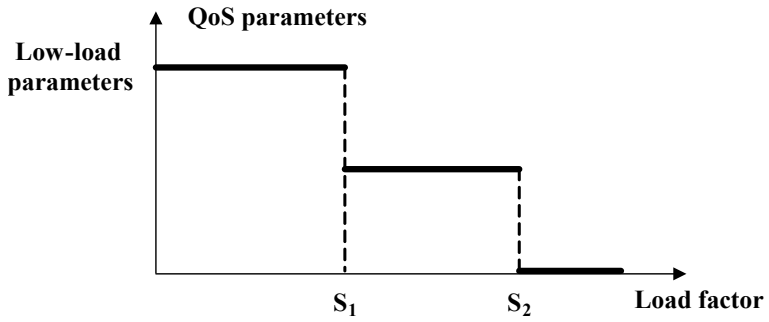


Fig. 7 QoS parameters reduction using the Staircase policy

Usually, d_{min} and d_{max} have different values. For instance, at low load factor, $d_{min}(x)$ and $d_{max}(x)$ are equal to 1 and 1.5 Mb/s, respectively. They are respectively reduced to 0.5 and 1 Mb/s as S_1 is reached, and are both set to zero when S_2 is exceeded. Furthermore, it is worth noting that the different serving RATs can have different S_1 and S_2 values.

5.2 Slope Tuning Policy

As radio access technologies are progressively loaded, the Slope policy gradually tunes QoS parameters as a function of the load factor (cf. Fig. 8). When RAT x load factor is low, the highest $d_{min}(x)$ and $d_{max}(x)$ values are signaled. Yet, when S_1 is reached, QoS parameters are linearly and separately reduced down to zero. The slope helps to better respond to traffic load fluctuations.

As QoS parameters are dynamically modulated, arriving mobiles are pushed to the less loaded RATs, enhancing long-term network performance. However, using both policies, the challenge is to properly set S_1 and S_2. In the same load conditions, QoS parameters to signal strongly depend on tuning threshold values. In other words, for a given load factor, different d_{min} and d_{max} can be provided depending on S_1 and S_2, leading to different user decisions. The impact of S_1 and S_2, on network and user utilities, are further discussed in this paper.

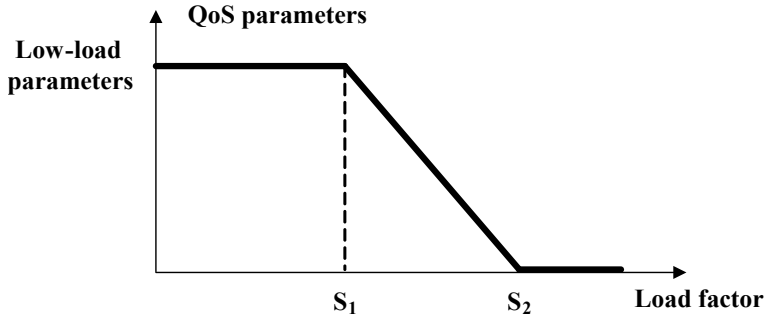


Fig. 8 QoS parameters reduction using the Slope policy

6 Performance evaluation

For simulation, we have developed a Matlab-based event-driven simulator adapted for heterogeneous wireless networks. RAT selection functionality is implemented according to our hybrid decision framework proposed in section 3.

N_T generic OFDM(A)-based radio access technologies are considered. RAT x capacity is fixed to C^x. The radio resource is divided into N_{RU}^x resource units (*i.e.,* OFDM symbols or OFDMA slots). In the time domain, transmissions are further organized into radio frames of length T^x.

At each scheduling epoch, resource units are allocated to individual users based on their priority and current needs (*i.e.,* amount of traffic waiting for transmission). Before any scheduling is applied, the minimum guaranteed RUs are directly granted. The Weighted Fair Queuing (WFQ) is then adopted to share out remaining resources. Grants are however limited to n_{max}. Session weights in WFQ schedulers are based on the cost they pay for one unit of traffic. Residual resources are afterwards equitably distributed (*i.e.,* according to the Round Robin service discipline).

Because network information may be dynamically tuned typically as a function of the load conditions, all mobiles do not necessarily perceive the same cost and QoS parameters, leading to different decision-makings. We therefore suppose that mobiles arrive sequentially. The total number of users is limited to N_{total}; it sets the traffic load. Their sojourn time is considered to be much greater in comparison with the simulation time $T_{simulation}$. Consequently, the network dynamics will progressively slow down until a pseudo-stationary regime is attained, where all measurements are performed. To improve the statistical significance of the results, simulations are repeated 500 times and performance metrics are averaged.

After they arrive, mobiles randomly select a user profile (cf. Table 1). As a matter of fact, they initiate either an inelastic, a streaming, or an elastic session, and determine their cost tolerance parameter λ and the weights $w_{d_{min}}$, $w_{d_{max}}$, and w_{cost} they apply to normalized decision criteria.

Profile No.	Traffic class	λ	$w_{d_{min}}$	$w_{d_{max}}$	w_{cost}
1	Inelastic	60	0.7	0	0.3
2	Streaming	60	14/30	7/30	0.3
3	Elastic	60	0	0.7	0.3
4	Inelastic	25	0.3	0	0.7
5	Streaming	25	0.2	0.1	0.7
6	Elastic	25	0	0.3	0.7

Table 1 User profiles

In Table 1, the weights of the decision criteria are normalized such that they sum up to 1 for each user profile. Further, mobiles decode current cost and QoS information, evaluate their expected satisfaction levels, and rank the different alternatives. The needs of inelastic and streaming sessions are respectively expressed as fixed (*i.e.,* R_f) and average long-term throughput (*i.e.,* R_{av}). We assume that the set of possible throughput demands is given by $D = \{0.5, 1, 1.5, 2\}$ Mb/s.

Inelastic and streaming traffic is packetized into small units of fixed length $L^c, c \in \{I, S\}$. Inelastic sessions generate packets according to a deterministic distribution, whereas streaming sessions generate packets according to a Poisson process. These packets are segmented into blocks sized to fit one RU. In our work, we fix delay constraints for the latter session types. A maximum delay requirement of $\Delta^c, c \in \{I, S\}$ is fixed. Since resources are limited, some packets may miss their deadline. They will be dropped as they are no longer useful.

Furthermore, the needs of elastic sessions are expressed as comfort throughput (*i.e.,* R_c). We suppose that the set of possible comfort throughputs is given by $C = \{0.75, 1.25\}$ Mb/s. Inelastic and streaming sessions uniformly choose one of the possible throughput demands regardless of the user cost tolerance parameter. Yet, we assume in the following that the comfort throughput of elastic sessions is related to the user willingness to pay, and thus imposed by the user profile.

To provide a detailed performance evaluation, three simulation scenarios are considered. In the first one, QoS information is investigated: we study the performance improvement achieved by providing differentiated service classes and minimum bandwidth guarantees to mobile users, regardless of future network load conditions. The second scenario compares our satisfaction-based multi-criteria decision-making method with other existing algorithms, namely SAW and TOPSIS. In the third scenario, we illustrate the gain from using our tuning policies in comparison with a static one.

6.1 Scenario 1: QoS information

In this first scenario, we are interested in the performance improvement achieved by providing differentiated service classes and minimum bandwidth guarantees to mobile users, regardless of future network load conditions.

We consider a realistic and cost-effective deployment, where N_T RATs are co-localized: the same base station site is used leading to cells overlapping. For the sake of simplicity, all users are assumed to belong to the same zone Z_k: they all have the same modulation and coding schemes, and thus exploit in the same manner their allocated grants. General simulation parameters are listed in table 2.

Parameters	Values
N_T	3
$C^x, x = 1, ..., N_T$	35 Mb/s
$N^x_{RU}, x = 1, ..., N_T$	700
$T^x, x = 1, ..., N_T$	10 ms
$T_{simulation}$	300 s
$L^c, c = I, S$	125 bytes
$\Delta^c, c = I, S$	100 ms

Table 2 Simulation parameters for the first and second scenarios

To evaluate long-term network performance, five major key performance indicators are defined: throughput, mean waiting delay and packet loss rate (for inelastic and streaming sessions), user-perceived satisfaction and operator gain. In our work, the waiting delay represents the time that a packet spends in the queue before being transmitted.

6.1.1 Service differentiation

To examine the impact of service differentiation on global network performance, we compare the following two situations:

- *Situation 1: Differentiated services network.* Access technologies provide differentiated service classes, namely, Premium, Regular and Economic. They differ in their QoS and cost parameters.
  A QoS-aware pricing scheme should be adopted: mobiles are charged based on their priority. Otherwise, all sessions would select the premium service class, and our differentiated services model would lose its interest.
- *Situation 2: Mono-service network.* Access technologies provide a unique service class, namely Regular plus.

Initial QoS and cost parameters, as perceived by mobile users, are depicted in Table 3. They are assumed fixed and do not change as the RAT load changes, except when the RAT is no longer able to guarantee to future arrivals the initial QoS parameters.

Service class	d_{min} (Mb/s)	d_{max} (Mb/s)	$Cost$ (unit/kB)
Premium	1.5	2	6
Regular	1	1.5	4
Economic	0.5	1	2
Regular Plus	1	2	4

Table 3 Static QoS and cost parameters

While inelastic sessions require inflexible QoS parameters, selection decisions must satisfy their fixed throughput demand. When the RAT is highly loaded, the resource scheduler is no more able to provide them with more than their minimum guaranteed throughputs, thus eventually leading to performance degradation. So as to enhance their QoS level, typically at high traffic load, mobiles should be provided with high enough bandwidth guarantees, or equivalently with high enough priority. Regardless of the user profile, selection decisions, when differentiated services are provided, are reported in table 4.

Throughput Needs (Mb/s)	**0.5**	**1**	**1.5**	**2**
Premium			✓	✓
Regular		✓		
Economic	✓			

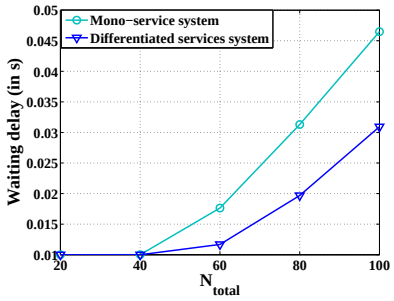
Table 4 Satisfaction-based decisions for inelastic sessions

Figures 9(a) and 9(b) respectively show the mean waiting delay and the packet drop probability as a function of the total number of arrivals. When differentiated services are provided, throughput-intensive sessions select the Premium service class with the highest priority, thus leading to a shorter delay, a lower drop probability and subsequently a better overall QoS level.

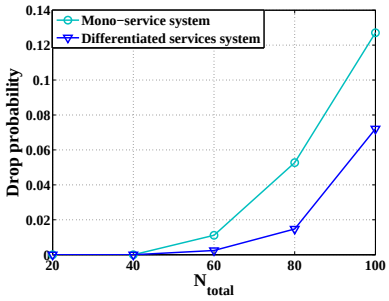
We depict in Fig. 9(c) the average user satisfaction. We notice that, at low traffic load, user satisfaction is higher when a unique service class is provided: the Regular plus service class fulfills strict QoS requirements, while charging mobile users on average with lower cost. Yet, when the network gets loaded, throughput-intensive sessions see their performance degraded: the Regular plus service class is no more able to meet their inflexible throughput demands, thus strongly decreasing the average perceived satisfaction. However, when differentiated services are provided, throughput-intensive sessions always opt for the Premium service class, thus enjoying higher bandwidth guarantees, leading to a larger overall satisfaction.

Furthermore, since streaming sessions are fairly flexible, mobiles may be less restrictive in their choices. Based on their preferences, users may actually look for fair enough content quality (average long-term throughput), high content quality (higher throughput) or even poor content quality (lower throughput). Selection decisions are put forward in tables 5 and 6.

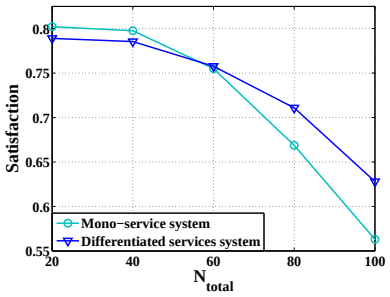
The mean waiting delay and the packet drop probability are respectively illustrated in Fig. 10(a) and 10(b). When differentiated services are provided,



(a) Mean packet delay



(b) Packet drop probability



(c) User satisfaction

Fig. 9 Inelastic sessions performance

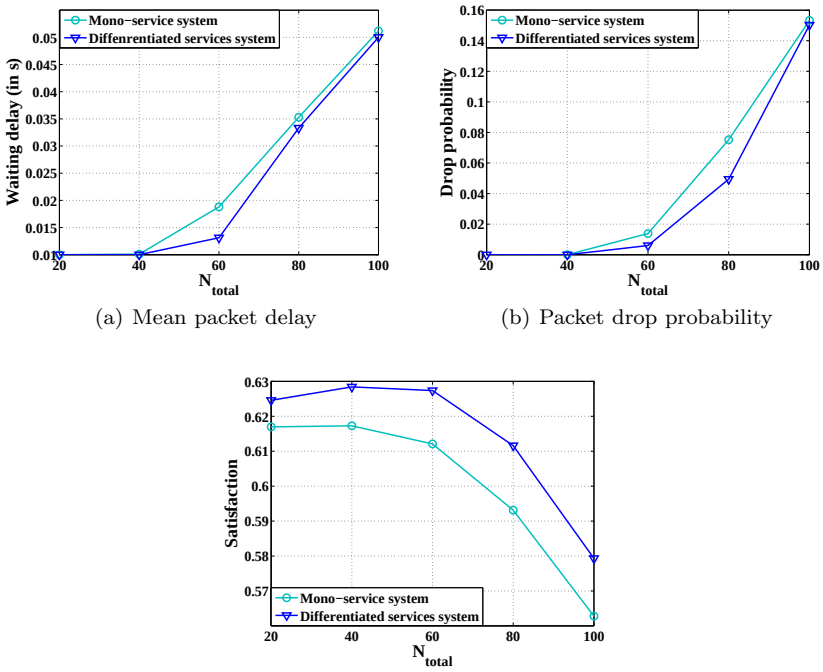
Throughput Needs (Mb/s)	**0.5**	**1**	**1.5**	**2**
Premium		✓	✓	✓
Regular	✓			
Economic				

Table 5 Satisfaction-based decisions for streaming sessions - users are ready to pay for better performances

Throughput Needs (Mb/s)	**0.5**	**1**	**1.5**	**2**
Premium				✓
Regular			✓	
Economic	✓	✓		

Table 6 Satisfaction-based decisions for streaming sessions - users seek to save up money

better performances are mainly observed at medium traffic load: demanding sessions could be provided with higher bandwidth guarantees (*i.e.,* with the Premium service class), and even low-priority sessions are granted more than their minimum guaranteed throughputs. However, when the network gets loaded, mobiles that seek to save up money (*i.e.,* users with profile no. 5), and thus have on average lower bandwidth guarantees, suffer from poor performances. On the other side, mobiles that are ready to pay (*i.e.,* users with



(a) Mean packet delay

(b) Packet drop probability

(c) User satisfaction

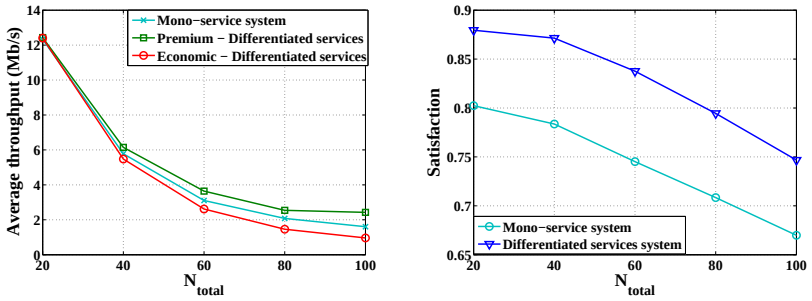
Fig. 10 Streaming sessions performance

profile no. 2) are always provided with high enough bandwidth guarantees, and consequently have better QoS than when a unique service class is offered. Therefore, at high traffic load, performances are on average very close: streaming sessions that are ready to pay offset the performance degradation of sessions that seek to save up money.

Besides, user satisfaction is constantly higher when differentiated services are provided (Fig. 10(c)). In contrast to inelastic sessions, users that seek to save up money sacrifice within limits their service quality (*i.e.,* select a cheaper service class) leading to a higher overall satisfaction, typically at low traffic load.

Because elastic sessions have no QoS needs, selection decisions exclusively depend on user preferences. Mobiles that are ready to pay (*i.e.,* users with profile no. 3) select the Premium service class, and then enjoy the highest throughput. However, those who seek to save up money (*i.e.,* users with profile no. 6) select the Economic one and thus have the lowest throughput. On the other hand, when a unique service class is provided, all sessions have similar priorities, leading to similar throughputs, as shown in Fig. 11(a).

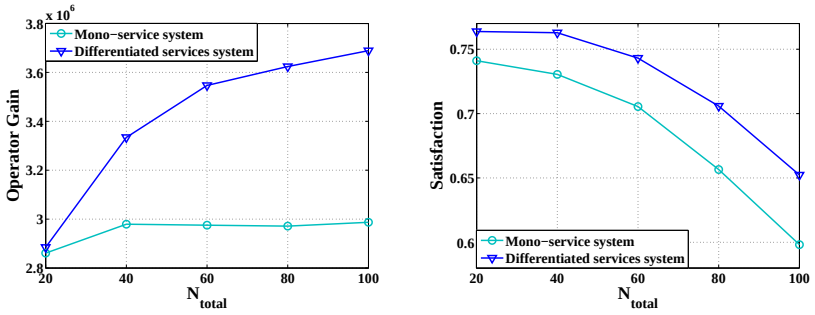
Since they are associated with the service class that best meets their preferences, elastic sessions have significantly higher satisfaction (Fig. 11(b)), when differentiated services are provided.



(a) Average throughput (b) User satisfaction

Fig. 11 Elastic sessions performance

In addition, as depicted in Fig. 12(a), operator gain is maximized, when differentiated services are proposed. However, although mobiles pay on average more, they have a significantly higher satisfaction (Fig. 12(b)). Actually, in a differentiated services network, users avoid undersized and oversized decisions, and are usually associated with the service class that best meets their needs and preferences.

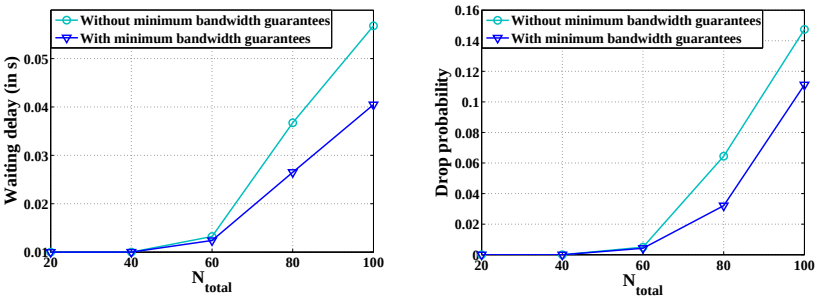


(a) Operator gain (b) User-perceived satisfaction

Fig. 12 Global network performance

6.1.2 Bandwidth guarantees

We also discuss the impact of bandwidth guarantees on real-time sessions performance: when real-time sessions (*i.e.,* inelastic and streaming sessions) are provided with minimum bandwidth guarantees (*i.e.,* $d_{min} \neq 0$) regardless of future network load, they have a shorter delay (Fig. 13(a)), a lower drop probability (Fig. 13(b)) and thus a better QoS level. Actually, real-time sessions will be always provided with, at least, their minimum guaranteed RUs, thus enhancing their performance typically when RATs get loaded.



(a) Mean packet delay (b) Packet drop probability

Fig. 13 Real-time sessions performance

6.2 Scenario 2: Multi-criteria decision-making methods

In this second scenario, we compare our Satisfaction-Based (SB) multi-criteria decision-making method to the well-known SAW [35] and TOPSIS [13,5] algorithms. As in the first scenario, hybrid cells include N_T co-localized RATs. Since we mainly focus on the decision-makings, and for the sake of simplicity, all mobiles are supposed to belong to the same zone Z_k. Thus, they are assumed to have the same peak rate. General simulation parameters are depicted in table 2.

Each RAT proposes three different service classes, namely Premium, Regular and Economic. QoS and cost parameters, as perceived by mobile users, are depicted in Table 7. Once again, they are supposed fixed and do not change as the RAT load changes, except when the RAT is no longer able to guarantee to future arrivals the initial QoS parameters.

Service class	d_{min} (Mb/s)	d_{max} (Mb/s)	*cost* (unit/kB)
Premium	1.5	2	6
Regular	1	1.5	4
Economic	0.5	1	2

Table 7 Static QoS and cost parameters

Before we discuss simulation results, let us recall the SAW and TOPSIS methods. When normalizing decision criteria $d_{min}(a)$, $d_{max}(a)$, and $cost(a)$, SAW and TOPSIS ignore user needs (*i.e.,* traffic class, throughput demand, cost tolerance), and exclusively depend on available alternatives. We note A the set of available alternatives and $\tilde{a}$ any element that belongs to A.

6.2.1 Simple Additive Weighting (SAW)

For alternative a, the normalizing functions regardless of the session traffic class c are:

$$\hat{d'}(a) = \frac{d'(a).g(M,C)}{\max_{\tilde{a}\in A} d'(\tilde{a}).g(M,C)} \tag{6}$$

where $d' = \{d_{min}, d_{max}\}$, and

$$\widehat{cost}(a) = \frac{\min_{\tilde{a}\in A} cost(\tilde{a})}{cost(a)} \tag{7}$$

The utility function of a class c session for alternative a is defined by :

$$U^c(a) = w^c_{d_{min}}.\hat{d}_{min}(a) + w^c_{d_{max}}.\hat{d}_{max}(a) + w^c_{cost}.\widehat{cost}(a)$$

Mobiles actually select the alternative with the highest score (*i.e.,* utility function).

6.2.2 Technique for Order Preference by Similarity to Ideal Solution (TOPSIS)

For alternative a, the normalizing functions regardless of the session traffic class c are:

$$\hat{d'}(a) = \frac{d'(a).g(M,C)}{\sqrt{\sum_{\tilde{a}\in A} (d'(\tilde{a}).g(M,C))^2}} \tag{8}$$

where $d' = \{d_{min}, d_{max}\}$, and

$$\widehat{cost}(a) = \frac{cost(a)}{\sqrt{\sum_{\tilde{a}\in A} (cost(\tilde{a}))^2}} \tag{9}$$

The positive and the negative ideal solutions, respectively denoted by a^+ and a^-, are then determined as follows:

$$a^+ = (d^+_{min}, d^+_{max}, cost^+) = (\max_{\tilde{a}\in A} \hat{d}_{min}(\tilde{a}), \max_{\tilde{a}\in A} \hat{d}_{max}(\tilde{a}), \min_{\tilde{a}\in A} \widehat{cost}(\tilde{a})) \tag{10}$$

$$a^- = (d^-_{min}, d^-_{max}, cost^-) = (\min_{\tilde{a}\in A} \hat{d}_{min}(\tilde{a}), \min_{\tilde{a}\in A} \hat{d}_{max}(\tilde{a}), \max_{\tilde{a}\in A} \widehat{cost}(\tilde{a})) \tag{11}$$

These ideal solutions do not necessarily exist: a^+ and a^- are defined as virtual alternatives with respectively the best and the worst decision criteria values.

The distance of alternative a from the positive ideal and the negative ideal solution, respectively denoted by $S^+(a)$ $S^-(a)$, are furthermore computed as:

$$S^+(a) = \sqrt{[w^c_{d_{min}}(\hat{d}_{min}(a) - d^+_{min})]^2 + [w^c_{d_{max}}(\hat{d}_{max}(a) - d^+_{max})]^2 + [w^c_{cost}(\widehat{cost}(a) - cost^+)]^2} \tag{12}$$

$$S^{-}(a) = \sqrt{[w^{c}_{d_{min}}(\hat{d}_{min}(a) - d^{-}_{min})]^2 + [w^{c}_{d_{max}}(\hat{d}_{max}(a) - d^{-}_{max})]^2 + [w^{c}_{cost}(\widehat{cost}(a) - cost^{-})]^2} \tag{13}$$

The relative closeness (*i.e.,* utility function) is however defined as:

$$C(a) = \frac{S^{-}(a)}{S^{-}(a) + S^{+}(a)} \tag{14}$$

Mobiles select the alternative with the shortest distance from the positive ideal solution and the farthest distance from the negative ideal solution, or equivalently the alternative with the highest relative closeness.

Because they ignore user needs, SAW and TOPSIS often lead to undersized and oversized decisions. When selections are independent of session throughput demands, users with a demand of 2 Mb/s make the exactly same decisions as those with a demand of 0.5 Mb/s. As a matter of fact, their decisions exclusively depend on user preferences (*i.e.,* weights of the decision criteria), as well as on the available alternatives. On the one hand, when users seek to save up money, they always opt for the Economic service class (*i.e.,* their best trade-off between QoS and cost parameters). As a consequence, the performance of throughput-intensive sessions are dramatically degraded. On the other hand, when they are ready to pay for better performances, they always select the Premium service class. Consequently, sessions with relatively low throughput demand will uselessly pay more: premium guarantees may not improve their performance in comparison with regular or economic ones.

Yet, our proposed Satisfaction-Based (SB) algorithm provides the best performance for the best cost. On the one hand, when session needs are stringent and inflexible, a high enough priority service class is selected, thus enhancing user-perceived performance. On the other hand, when higher bandwidth guarantees do not improve session performance, SB leads to a low enough priority service class, thus charging mobile users with lower cost. So as to make the comparison more fair, *enhanced SAW and TOPSIS* are used: they only explore feasible alternatives. When their throughput demand is greater than the provided d_{max}, the alternative opted for is considered to be infeasible and thus rejected. This will prevent SAW and TOPSIS from making some undersized decisions. However, as discussed in the following paragraph, our proposed method continues to outperform them.

6.2.3 Comparison results

So as to enhance network performance, and as stated above, enhanced SAW and TOPSIS only explore feasible alternatives. Yet, they continue to lead to some undersized, but mostly oversized alternatives. For inelastic sessions, selection decisions, according to the different multi-criteria decision-making methods, are reported in tables 8 and 9.

When users are ready to pay for better performances (*i.e.,* users with profile no. 1), SAW and TOPSIS always single out the Premium service class.

Intuitively, and since inelastic session needs are fixed, this decision is oversized for 0.5 and 1 Mb/s sessions. As SB respectively opts for the Economic and the Regular service classes, QoS requirements are always perfectly satisfied, while charging mobile users with lower cost.

Decision Method	**SAW/TOPSIS**				**SB**			
Session Needs (Mb/s)	**0.5**	**1**	**1.5**	**2**	**0.5**	**1**	**1.5**	**2**
Premium	✓	✓	✓	✓			✓	✓
Regular						✓		
Economic					✓			

Table 8 Inelastic sessions - users are ready to pay for better performances

Also, when users seek to save up money (*i.e.,* users with profile no. 4), enhanced SAW and TOPSIS lead to the Economic service class for 1 Mb/s sessions, and to the Regular service class for 1.5 Mb/s sessions. These decisions are undersized. When the RAT is highly loaded, fixed QoS requirements are not satisfied, thus dramatically degrading session performances.

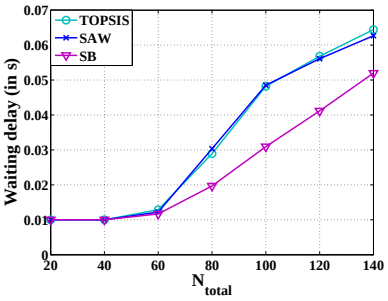
Decision Method	**SAW/TOPSIS**				**SB**			
Session Needs (Mb/s)	**0.5**	**1**	**1.5**	**2**	**0.5**	**1**	**1.5**	**2**
Premium				✓			✓	✓
Regular			✓			✓		
Economic	✓	✓			✓			

Table 9 Inelastic sessions - users seek to save up money

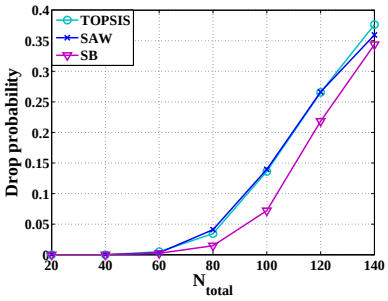
Figures 14(a) and 14(b) respectively show the mean waiting delay and the packet drop probability as a function of the total number of arrivals. Since it avoids undersized decisions, SB provides a shorter delay, a lower drop probability and subsequently a better overall QoS level.

We depict in Fig. 14(c) the average user satisfaction. We notice that, at low traffic load, enhanced SAW and TOPSIS provide higher satisfaction. First, undersized decisions are able to fulfill strict QoS requirements, while charging mobile users less. Second, although oversized decisions decrease user satisfaction, the reduction is not significant enough to offset the impact of undersized decisions. In other words, at low traffic load, undersized decisions considerably increase user satisfaction, because the corresponding users seek to save up money. Their QoS needs are perfectly met, while paying less. However, oversized decisions do not significantly decrease user satisfaction, because users in question are originally ready to pay. We further note that, when traffic load is moderate, SB brings the largest satisfaction since it always meets the strict QoS requirements. In fact, using SAW and TOPSIS, undersized decisions are no more able to meet user needs when traffic load is relatively high.

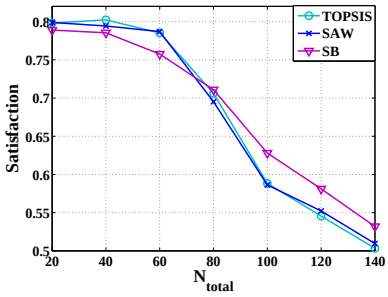
For streaming sessions, selection decisions are put forward in tables 10 and 11. When users are ready to pay for better performances, for 0.5 Mb/s



(a) Mean packet delay



(b) Packet drop probability



(c) User satisfaction

Fig. 14 Inelastic sessions performance

Decision Method	**SAW/TOPSIS**					**SB**		
Session Needs (Mb/s)	**0.5**	**1**	**1.5**	**2**	**0.5**	**1**	**1.5**	**2**
Premium	✓	✓	✓	✓		✓	✓	✓
Regular					✓			
Economic								

Table 10 Streaming sessions - users are ready to pay for better performances

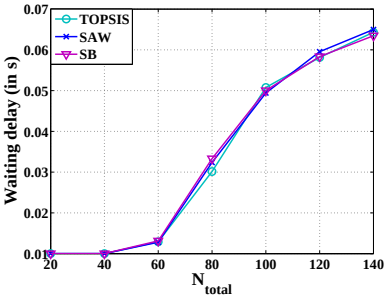
Decision Method	**SAW/TOPSIS/SB**			
Session Needs (Mb/s)	**0.5**	**1**	**1.5**	**2**
Premium				✓
Regular			✓	
Economic	✓	✓		

Table 11 Streaming sessions - users seek to save up money

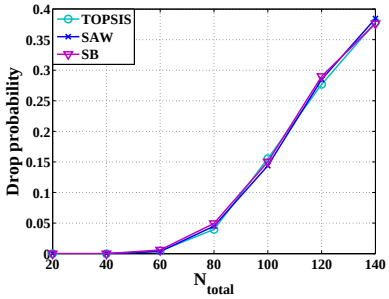
sessions, SAW and TOPSIS lead to the Premium service class and SB to the Regular one. SAW and TOPSIS decisions are oversized. The Regular service class actually provides users with twice their average long-term throughput.

The mean waiting delay and the packet drop probability are respectively depicted in Fig. 15(a) and 15(b). Since all methods provide the same QoS level, the Premium service class proves to be oversized for 0.5 Mb/s sessions.

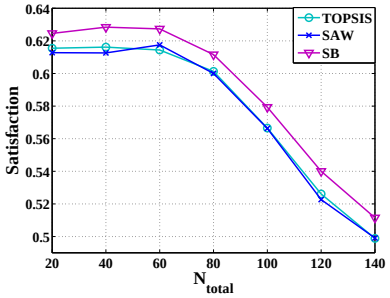
In comparison with SB, no performance improvement is observed. Therefore, on average, SB charges less and carries out higher user satisfaction (Fig. 15(c)).



(a) Mean packet delay



(b) Packet drop probability



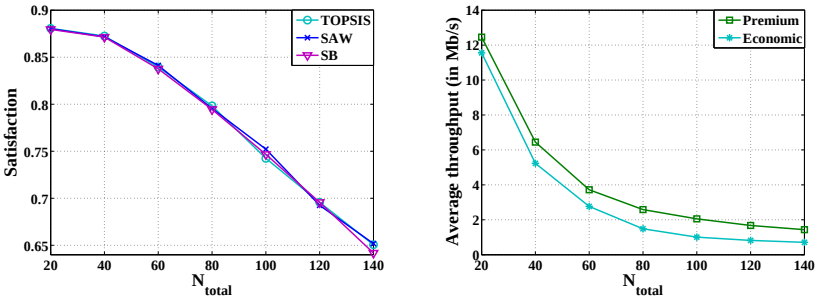
(c) User satisfaction

Fig. 15 Streaming sessions performance

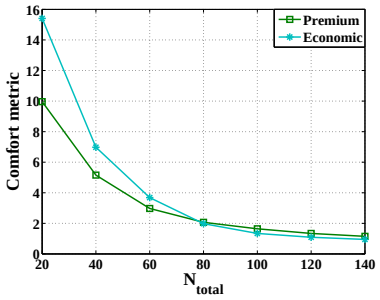
Because elastic sessions accommodate with available bandwidth, undersized and oversized decisions do not technically exist. When SB takes into account user comfort throughput, it may theoretically reach different solutions from SAW and TOPSIS. Yet, given our simulation model and parameters, they practically all lead to the same decisions, providing the same user satisfaction (cf. Fig. 16(a)).

When users are ready to pay for better performances (*i.e.,* users with profile no. 3), they systematically select the Premium service class. Nevertheless, when they seek to save up money (*i.e.,* users with profile no. 6), they choose the Economic one. As illustrated in Fig. 16(b), Premium sessions enjoy higher throughputs than Economic ones.

The comfort metric is defined as the ratio of the perceived throughput to the comfort one. Although Premium sessions have higher throughputs, their comfort metric is similar to the Economic ones except at low traffic load (cf. Fig. 16(c)). Thereby, our solution ensures fairness with respect to different comfort throughputs.



(a) User satisfaction (b) Mean perceived throughput



(c) Mean comfort metric

Fig. 16 Elastic sessions performance

When a RAT is no longer able to guarantee to future arrivals the initial QoS parameters, network information is modified. As they have lower bandwidth guarantees for the same initial monetary cost, new arrivals are considered to be disadvantaged. We depict in Fig. 17 the Disadvantaged Sessions Rate, denoted by DSR and defined as the number of disadvantaged sessions over the total number of on-going sessions. Since it avoids oversized decisions, SB brings the lowest DSR. At high traffic load, higher QoS guarantees are provided respectively with SB, SAW and TOPSIS.

To wrap up, SB avoids undersized decisions, best meets QoS requirements and brings the best performances. By eliminating infeasible alternatives, enhanced SAW and TOPSIS bring similar performances as SB, for streaming and elastic sessions. However, SB considerably outperforms them for inelastic sessions, where QoS requirements are stringent and inflexible.

Also, by evading oversized decisions typically for inelastic and streaming sessions, SB charges on average less than enhanced SAW and TOPSIS. Thereby, SB leads to better performances, lower cost and therefore higher user satisfaction.

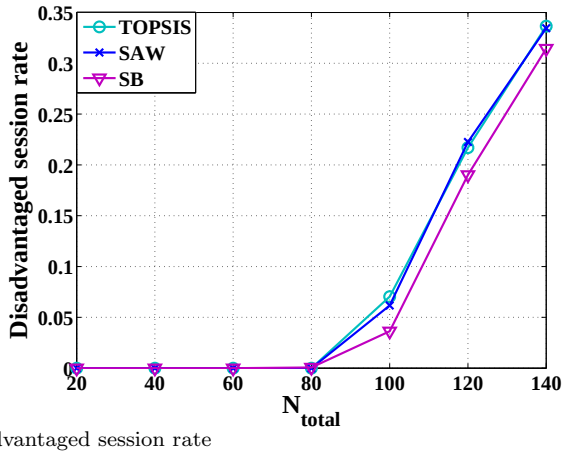


Fig. 17 Disadvantaged session rate

6.3 Scenario 3: Tuning policies

In this third scenario, we illustrate the gain from using our tuning policies in comparison with a static one. When a RAT dominates all the others (*i.e.,* provides higher QoS parameters for the same cost or the same QoS parameters for a lower cost), QoS information are either modulated as a function of the load conditions using the staircase or the slope tuning policies, or maintained fixed leading to performance inefficiency. General simulation parameters are however listed in table 12.

Parameters	Values
N_T	2
$C^x, x = 1, ..., N_T$	70 Mb/s
$N^x_{RU}, x = 1, ..., N_T$	700
$T^x, x = 1, ..., N_T$	10 ms
$T_{simulation}$	300 s
$L^c, c = I, S$	125 byte
$\Delta^c, c = I, S$	100 ms

Table 12 Simulation parameters for the third scenario

Each RAT is assumed to propose three different service classes, namely Premium, Regular and Economic. All RATs are supposed to initially signal the same QoS and cost parameters listed in table 13.

We further assume that mobiles randomly select a set of modulation and coding gains. These multiplicative factors reflect the user radio conditions in the different technologies, and are supposed to remain constant in time. Two sets of gains are considered and reported in Table 14. They typically illustrate the network topology of Fig. 18.

Service class	d_{min} (Mb/s)	d_{max} (Mb/s)	$Cost$ (unit/kB)
Premium	1	1.35	6
Regular	0.7	1	4
Economic	0.35	0.7	2

Table 13 Initial QoS and cost parameters

Set No.	RAT 1	RAT 2
1	1.5	1.5
2	2	1

Table 14 Modulation and Coding gain

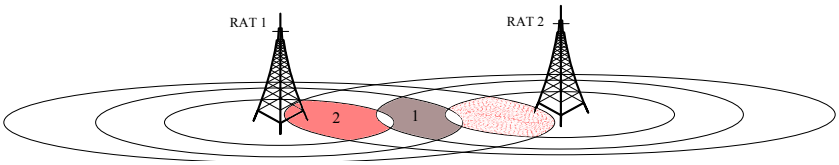


Fig. 18 Scenario 3: a possible network topology

When the two access technologies provide the same QoS parameters, users that are associated with set no. 2 would select RAT 1. They expect to have better radio conditions, and thus to perceive higher throughputs in RAT 1. All other alternatives, proposed by RAT 2, are subsequently dominated. Also, users that are associated with set no. 1 randomly join their RAT, since they expect to perceive similar throughputs in the two available RATs. This situation leads to unevenly distributed traffic load. However, when network information is dynamically modulated according to the staircase or to the slope tuning policies, QoS parameters are changed in a way to drive future arrivals to the less loaded RAT: loaded technologies provide lower QoS parameters, and thus push future users to less loaded technologies. When staircase policy is adopted, reduced QoS parameters are presented in Table 15.

Service class	d_{min} (Mb/s)	d_{max} (Mb/s)
Premium	0.5	0.7
Regular	0.35	0.5
Economic	0.2	0.5

Table 15 Reduced QoS parameters (staircase policy)

Other scenarios may also lead to unevenly distributed traffic load. For instance, when mobiles have the same modulation and coding schemes, a technology is preferred if it initially broadcasts higher QoS parameters for the same cost, or the same QoS parameters for a lower cost. While static information absolutely leads to performance inefficiency, dynamic tuning helps to better distribute mobile users over the available RATs, and thus to efficiently utilize radio resources.

When using the staircase or the slope tuning policies, we assume that S_1 and S_2 are respectively set to 0.5 and 0.9 times the RAT capacity. Before S_1, the network provides constant QoS parameters. After S_2, QoS incentives are no longer provided to future arrivals: the network keeps a margin of about 10% of the RAT capacity to provide on-going sessions with more than their minimum guaranteed throughputs. These parameters will be studied in subsection 7.1.

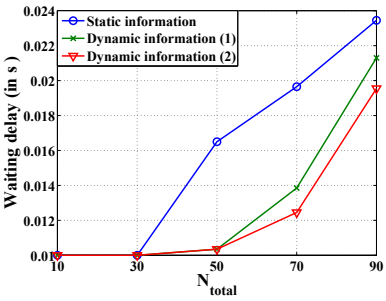
Results have shown the same trend for different simulation scenarios and parameters. Typically, we came to exactly the same conclusions with different modulation and coding gains, initial network information, network model parameters, tuning thresholds (*i.e.,* S_1 and S_2), and also when a unique service class is provided.

Because real-time (RT) sessions (*i.e.,* inelastic and streaming sessions) require tight delay constraints, access technologies should meet their throughput demands. However, users with a demand of 2 Mb/s may suffer: even the Premium guarantees may be lower than their throughput demand. When the RAT is highly loaded, the resource scheduler will not be able to provide them with more than their minimum guaranteed throughputs, thus leading to packet loss. So as to reduce the packet drop probability, we should avoid that a RAT gets overloaded long before the others. Load balancing should then be achieved.

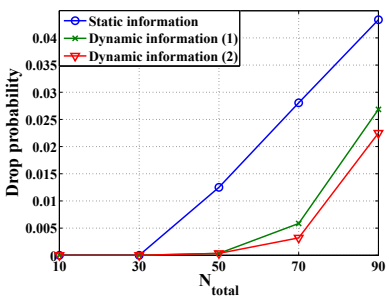
Figures 19(a) and 19(b) respectively show the mean waiting delay and the packet drop probability as a function of the total number of arrivals. When the slope intervention policy denoted as Dynamic information (2) is adopted, it best responds to traffic load fluctuations, and thus provides a shorter delay, a lower drop probability and subsequently a better overall QoS level. On the other hand, the staircase intervention policy denoted as Dynamic information (1) is disadvantageous when all technologies have exceeded their S_1: while load conditions are critical, RAT 1 is once again privileged until the operator guarantees exceed S_2 (*i.e.,* until RAT 1 no longer provides QoS guarantees to future arrivals). Yet, the performance of real-time sessions are always significantly enhanced in comparison with the static scenario, denoted as Static information.

Moreover, when sessions are better distributed over the two RATs, they will be allocated on average more RUs. Typically, when QoS parameters are tuned as a function of the load conditions, elastic sessions experience higher throughput and subsequently higher comfort metric, as shown in Fig. 19(c). However, at low traffic load (since tuning policies are not yet triggered) and at high traffic load (since all technologies become similarly occupied regardless of the tuning policy), performance enhancement is not that significant for elastic sessions.

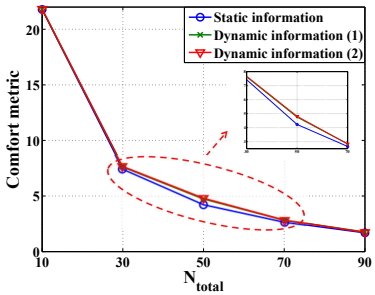
Furthermore, when tuning policies are triggered, QoS parameters are reduced. To benefit from the same initial bandwidth guarantees, mobile users may have to select a higher priority service class, and thus pay more. Also because fewer real-time packets are dropped (cf. Fig. 19(b)) and more elastic packets are served (cf. Fig. 19(c)), users consume on average a larger amount of traffic (Fig. 20(a)), and once again pay more. We illustrate in Fig. 20(b)



(a) Mean packet delay for RT sessions



(b) Mean packet drop probability for RT sessions



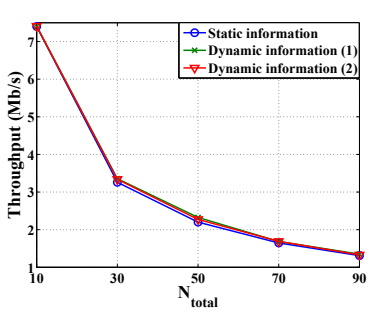
(c) Mean comfort metric for elastic sessions

Fig. 19 Measured QoS

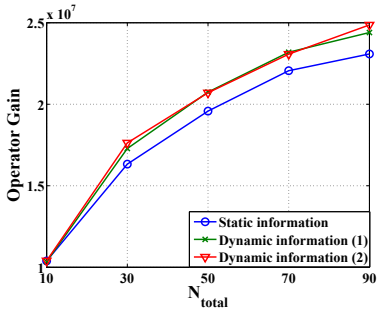
the average operator gain. When operators dynamically intervene, they gain more.

We depict in Fig. 20(c) the average user satisfaction. Although mobiles may pay more, we notice a higher satisfaction when tuning policies are implemented. Higher costs are then justified, since users benefit from significantly better performances. At low traffic load, tuning policies are not yet triggered. Equivalent performances, costs and subsequently satisfactions are intuitively observed. However, at very high traffic load, the performance gain over the static scheme begins to reduce; henceforth, it slightly offsets the cost considerations, leading to close user satisfaction.

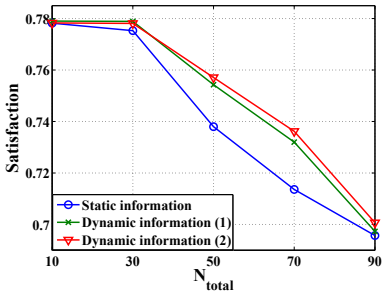
To conclude, in comparison with the static scheme, performance results show that our tuning policies enhance network performances, provide larger operator gain and higher user satisfaction. Since it best responds to traffic load fluctuations, the slope tuning policy has proved to be an efficient strategy that enhances resource utilization. In [11,12], we have formulated tuning policies as a Semi-Markov Decision Process (SMDP), and derived optimal solutions.



(a) Mean perceived throughput



(b) Operator gain



(c) User-perceived satisfaction

Fig. 20 Operator gain and global network performance

7 Comparison with different RAT selection schemes

In this section, we compare our hybrid approach with different RAT selection schemes including network-centric, hybrid and mobile-terminal-centric methods.

For illustration, we consider a heterogeneous wireless network composed of Mobile WiMAX and LTE RATs. They are supposed to utilize a channel bandwidth of 5 and 10 MHz respectively. Although our solution adapts to different deployment scenarios, here again, we focus on the more realistic and cost effective one where the two RATs base stations are co-localized. The intersection of their respective zones thus leads to N_Z heterogeneous zones.

For the sake of simplicity, the cell is assumed divided into two zones (*i.e.,* $N_Z = 2$). While users with good radio conditions (*i.e.,* in zone 1) are considered adopting the (64 - QAM, 3/4) modulation and coding scheme, users with bad radio conditions (*i.e.,* in zone 2) are supposed to employ the (16 - QAM, 1/2) one. Their peak rates are reported in Table 16.

Radio resources are allocated using fair time scheduling. Yet, when our hybrid method is employed, mobiles are first provided with their minimum guaranteed throughput given by d_{min}. Then, fair time scheduling is used to provide them with up to their maximum throughput given by d_{max}. Remaining

RAT	64-QAM: 3/4	16-QAM: 1/2
Mobile WiMAX (5 MHz)	16.6 Mb/s	7.4 Mb/s
LTE (10 MHz)	33.5 Mb/s	14.9 Mb/s

Table 16 Peak rates in Mobile WiMAX and LTE

resources are afterwards equitably shared (*i.e.,* after receiving their maximum throughput, all mobiles have the same priority leading to fair time scheduling).

Streaming and elastic sessions are individually considered. As in section 6, mobiles are randomly ready either to pay for better performances, or to sacrifice within limits their service quality seeking to save up money. When user decisions need to be evaluated, or typically when their perceived satisfaction is to be computed, a set of cost tolerance parameter and QoS and cost weights is used according to user preferences (cf. Table 17).

Set No.	λ	w_{QoS}	w_{cost}
1	60	0.7	0.3
2	45	0.3	0.7

Table 17 Cost tolerance parameter and QoS and cost weights

We assume that streaming sessions have an average long-term throughput of 1 Mb/s. So as to improve content quality, they can furthermore benefit from throughputs up to 1.5 Mb/s (*i.e.,* $R_{av} = 1$ Mb/s and $R_{max} = 1.5$ Mb/s). When our proposed hybrid approach is used, the cost tolerance parameter and the weights that are assigned to the decision criteria (*i.e.,* d_{min}, d_{max} and *cost*) are put forward in Table 18. When profile no. 1 is assigned to users that are ready to pay for better performances, profile no. 2 is attributed to those that seek to save up money.

Profile No.	λ	$w_{d_{min}}$	$w_{d_{max}}$	w_{cost}
1	60	14/30	7/30	0.3
2	45	0.2	0.1	0.7

Table 18 User profiles for streaming sessions

Besides, elastic sessions adapt to resource availability. Their needs are expressed as comfort throughput, denoted by R_c. As in section 6, we assume that R_c is related to the user willingness to pay, and thus imposed by the user profile (cf. Table 19). Typically, when users are ready to pay for better performances, they have a comfort throughput of 1.25 Mb/s. Yet, when they seek to save up money, they are content with a comfort throughput of 0.75 Mb/s.

When our hybrid method is employed, initial network parameters are reported in Table 20. As RATs get loaded, their signaled QoS parameters are linearly reduced down to zero (*i.e.,* dynamically tuned according to the slope

Profile No.	λ	$w_{d_{min}}$	$w_{d_{max}}$	w_{cost}	R_c (Mb/s)
1	60	0	0.7	0.3	1.25
2	45	0	0.3	0.7	0.75

Table 19 User profiles for elastic sessions

tuning policy). However, when different thresholds (*i.e.*, S_1 and S_2) are considered, different QoS parameters may be signaled for the same load conditions. This may lead to different decision-makings depending on S_1 and S_2. Consequently, and before we compare our hybrid approach with other selection schemes, let us study the effect of thresholds S_1 and S_2 on network performance and user satisfaction.

RAT	d_{min} (Mb/s)	d_{max} (Mb/s)	*cost* (unit/kB)
Mobile WiMAX	1	1.5	4
LTE	1.5	2	6

Table 20 Initial QoS and cost parameters

To evaluate selection decisions, network and user utilities are introduced. The network utility reflects operator objectives: it is defined as the total offered throughput. Furthermore, the user utility reflects the average user-perceived satisfaction: it depends on their needs and preferences, and thus takes into account both QoS and cost considerations.

7.1 Effect of S_1 and S_2

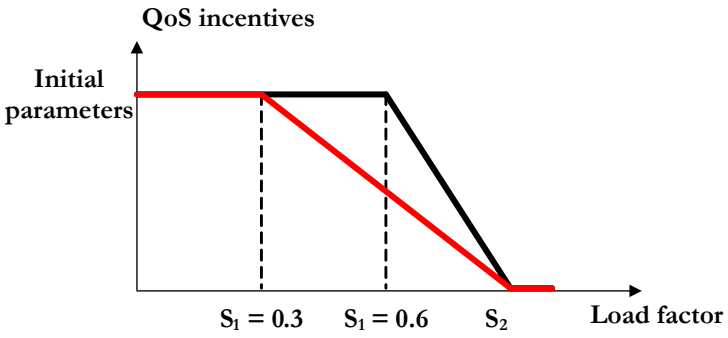


Fig. 21 S_1 effect on signaled QoS parameters

We illustrate in Fig. 21 the effect of S_1 on signaled QoS parameters. The lower S_1 is, the earlier d_{min} and d_{max} get reduced pushing more mobiles to less loaded RATs. Yet, the higher S_1 is, the steeper the slope is. The decay rate of the QoS parameters actually increases with S_1.

Moreover, figure 22 depicts the effect of S_2 on signaled QoS parameters. The lower S_2 is, the steeper the decrease of d_{min} and d_{max} is. Tuning becomes then more sensitive to load conditions. In other words, the lower S_2 is, the lower the QoS parameters are for the same load conditions, pushing more mobiles to less loaded RATs.

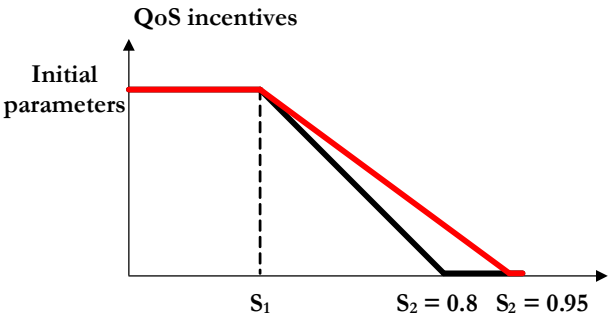


Fig. 22 S_2 effect on signaled QoS parameters

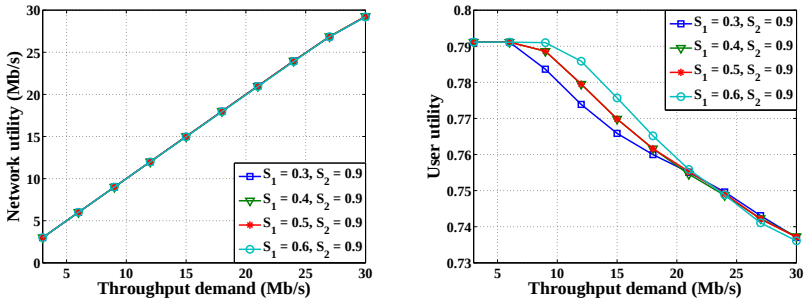
7.1.1 Streaming sessions

We first fix S_2 to 0.9 and vary S_1, so as to study its effect on network performance and user satisfaction.

We respectively show in Fig. 23(a) and 23(b) the network utility and the average user utility as a function of the total throughput demand. At very low traffic load, regardless of S_1, initial QoS parameters are broadcasted. Consequently, mobile WiMAX is generally preferred: it perfectly meets user QoS needs while charging them less. Only users, with bad radio conditions, that are ready to pay would select the LTE technology. Equivalent decision-makings are then observed for different S_1 values, leading to similar network and user utilities.

As WiMAX gets loaded, its broadcasted QoS parameters start to be reduced, pushing more arrivals to LTE. When different S_1 are examined, mobiles are differently distributed over the two RATs. Typically, when S_1 is fixed to 0.3, users are encouraged to join LTE much earlier than when S_1 is fixed to 0.6. As a result, at low and medium traffic load, the lower S_1 is, the more users join LTE and thus pay more. Similarly, the higher S_1 is, the more users continue to prefer mobile WiMAX competing for the same common resources. Yet, as shown in Fig. 23(a), mobiles can still achieve throughputs up to their R_{max} even for $S_1 = 0.6$.

Actually, since their throughput demands are limited, no performance difference is observable for streaming sessions depending on S_1 (cf. Fig. 23(a)). Even for $S_1 = 0.6$, at low and medium traffic load where more users join mobile WiMAX in comparison with other cases, the total offered throughput can still follow the throughput demand increase. Yet, since less users join LTE and pay

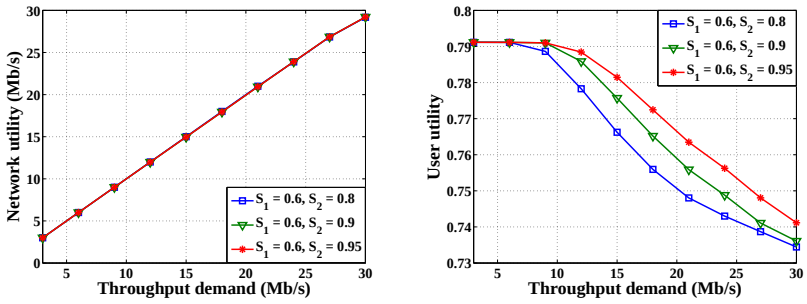


(a) Network utility (b) User utility

Fig. 23 S_1 effect on the performance of streaming sessions

more, users experience the highest satisfaction when $S_1 = 0.6$ (cf. Fig. 23(b)). However, at high traffic load, the proportion of users that are associated with LTE significantly increases for high S_1 values. While the QoS parameters signaled by the WiMAX technology are being roughly reduced (high decay rate), more and more mobiles join LTE. Therefore, in the long term, the average proportion of users that are connected to LTE becomes quite similar, regardless of S_1 values. This leads to fairly close network and user utilities at high traffic load.

Furthermore, we fix S_1 to 0.6 and vary S_2, so as to study its impact on network performance and user satisfaction. Following the same reasoning, the lower S_2 is, the more users are pushed to LTE. However, unlike for S_1, even when the total throughput demand is about 30 Mb/s, the proportion of users that are connected to LTE remains higher for lower S_2 values. As a matter of fact, the higher S_2 is, the longer can WiMAX provides attracting QoS guarantees for users. This leads to higher satisfaction (cf. Fig. 24(b)) seeing that users perceive similar performances (cf. Fig. 24(a)).



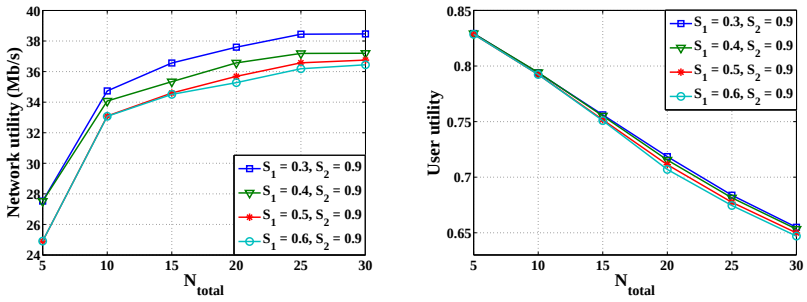
(a) Network utility (b) User utility

Fig. 24 S_2 effect on the performance of streaming sessions

7.1.2 Elastic sessions

Here again, we first fix S_2 to 0.9 and vary S_1 to study its impact on network performance and user satisfaction.

Figures 25(a) and 25(b) respectively illustrate the network utility and the average user utility as a function of the total number of users denoted by N_{total}. The lower S_1 is, the more efficiently mobiles are distributed over the two RATs. Typically, when $S_1 = 0.3$, broadcasted QoS parameters start to be reduced much earlier in comparison with other cases. As a result, more users particularly with good radio conditions join LTE, thus enhancing resource utilization.



(a) Network utility (b) User utility

Fig. 25 S_1 effect on the performance of elastic sessions

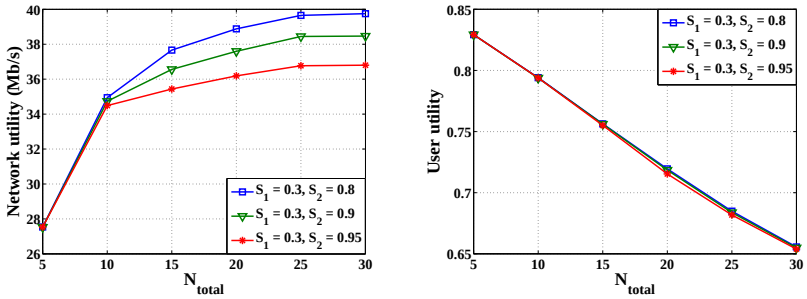
As a matter of fact, as tuning starts earlier, even mobiles with good radio conditions that are typically ready to pay (*i.e.,* having a comfort throughput of 1.25 Mb/s) start earlier to join LTE. Consequently, and since elastic sessions adapt to resource availability, the total offered throughput (*i.e.,* the network utility) is improved as shown in Fig. 25(a).

At low and medium traffic load, when S_1 is fixed to 0.3, more users particularly with good radio conditions join LTE in comparison with other cases. This better exploits LTE resources, enhancing network utility. Since less users are connected to WiMAX and more users including those with good radio conditions join LTE, users have on average better performances. Yet, as they pay on average more (more users are connected to LTE), users perceive close satisfaction regardless of S_1 values (cf. Fig 25(b)).

As N_{total} increases, the lower S_1 is, the higher is the average proportion of users with good radio conditions that are connected to LTE. This leads to continuously higher network utility. Thereby, and since in the long term the average proportion of users that are connected to LTE becomes close regardless of S_1 values, users perceive higher satisfaction for lower S_1 values.

Hereafter, we fix S_1 to 0.3 and vary S_2, so as to study its effect on network performance and user satisfaction. Following the same reasoning, the lower S_2

is, the more users particularly with good radio conditions join LTE leading to higher network utility (cf. Fig. 26(a)). On the other hand, as for streaming sessions, the higher S_2 is, the more users join WiMAX even for $N_{total} = 30$. As a consequence, for different S_2 values, cost considerations offset performance improvement leading to close user satisfaction (cf. Fig. 26(b)).



(a) Network utility (b) User utility

Fig. 26 S_2 effect on the performance of elastic sessions

7.2 Comparison results

In what follows, we compare our hybrid approach with six different RAT selection schemes:

- Peak rate maximization: Mobile users have no information on the global network state. Based on their radio conditions, they select the RAT that offers them the best peak rate.
- Instantaneous rate maximization: Mobiles are assumed to know the exact number of users that are connected to available technologies. Assuming that fair time scheduling is employed, they select the RAT that offers them the best throughput. Their estimated throughput in RAT x, $\overline{D^x}$, at the time of selection, is computed as:

$$\overline{D^x} = \frac{D^x}{1 + N^x} \tag{15}$$

  where D^x represents the user peak rate when connected to RAT x and N^x represents the number of users that are connected to RAT x at the time of selection.
- Satisfaction-based using peak rate (SB - PR): Using their peak rates, mobiles adopt the Satisfaction-based multi-criteria decision-making method to select their best RAT. In order to evaluate the different technologies, the provided QoS parameters, in Eq. 3 and 4, are replaced with the peak rate that mobiles can achieve when connected to these technologies.

- Satisfaction-based using instantaneous rate (SB - IR): Mobiles use the Satisfaction-based multi-criteria decision-making method to select the RAT that maximizes their expected utility. In Eq. 3 and 4, the provided QoS parameters are replaced with the estimated average throughput that mobiles can obtain.
- Exhaustive search: The network considers all possible associations involving all users. It finally selects the combination that optimizes its own utility. Actually, it assigns mobiles to either WiMAX or LTE technologies in a way to maximize the total offered throughput. This is known to be the optimal method with respect to operator objectives: it leads to the highest network utility.
- Our hybrid approach: The network periodically sends decisional information (*i.e.,* cost and QoS parameters) to assist mobile users in their decisions. A RAT is considered to be low-loaded when its load factor is below S_1. Initial d_{min} and d_{max} are then signaled (cf. Table 20). Yet, when its load factor exceeds S_2, a RAT is considered to be highly loaded, providing no QoS guarantees.

When using the peak rate maximization and the SB - PR methods, mobiles select their RAT without any network assistance. Decisions are then mobile-terminal-centric. However, when employing the instantaneous rate maximization and the SB - IR methods, load conditions signaled by the network assist mobile users in their decisions. The latter two methods are thus considered to be hybrid. Finally, when adopting the exhaustive search method, decisions are network-centric, since they are made by the network transparently to end-users.

Because in practice telecom operators will not reveal neither the exact numbers of users that are connected to their RATs nor the scheduling algorithm they adopt, the instantaneous rate maximization and the SB - IR methods are not realistic. Yet, they serve as a means to illustrate the gain from masking network load conditions and only signaling cost and some QoS parameters so as to enhance resource utilization.

7.2.1 Streaming sessions

Figures 27 and 28 respectively show the network utility and the average user utility as a function of the total throughput demand.

The network utility, defined as the total offered throughput, generally increases with the total throughput demand. Yet, when a RAT gets overloaded, its offered throughput stagnates and no longer increases with additional throughput demand.

When the SB - PR method is used, all users select the mobile WiMAX technology (*i.e.,* Mobile WiMAX is their best trade-off between cost and QoS decision criteria). Regardless of user preferences and radio conditions, mobile WiMAX is expected to provide mobile users with the highest utility. Since mobiles use their peak rate in estimating their utility, their decisions do not

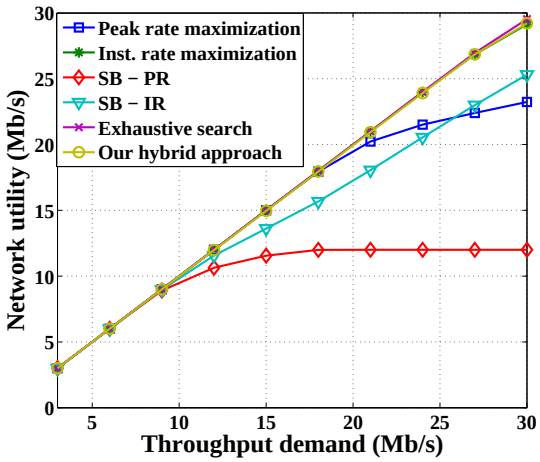


Fig. 27 Network utility: Streaming sessions scenario

depend on network load conditions. As a result, mobiles continue to select the WiMAX technology even when it gets overloaded.

At low traffic load, mobile WiMAX can meet user QoS needs, while charging them less. When users benefit from throughputs up to their R_{max} and pay less, they have the highest utility (*i.e.,* satisfaction). However, when WiMAX gets loaded, it becomes no longer able to fulfill user QoS needs. Typically, at medium and high traffic load, WiMAX becomes saturated leading to a significant decrease of the user-perceived throughput below R_{av} (cf. Fig. 27). As a consequence, user-perceived satisfaction will also dramatically decrease (cf. Fig. 28).

Furthermore, when the peak rate maximization method is adopted, all users select the LTE technology. Independently of their modulation and coding schemes, mobiles can achieve the best peak rate when connected to the LTE technology. Here again, their decisions do not change with network load conditions. As a consequence, at high traffic load, user-perceived throughput goes below R_{max}. Yet, it continues to be greater than R_{av}.

On the other hand, since LTE charges more than WiMAX does, mobile users experience the lowest satisfaction level at low traffic load. Actually, when all RAT selection schemes meet user QoS needs, the peak rate maximization method assigns all users to the LTE technology, thus charging them more. At high traffic load, because user-perceived throughput decreases, their experienced utility also diminishes.

Moreover, when the SB - IR method is employed, users combine their needs and preferences with network load conditions to select their best RAT. As a consequence, at low traffic load and regardless of their radio conditions, all users select the mobile WiMAX technology: their QoS needs are perfeclty met while paying less. This leads to the highest user-perceived utility, as in the

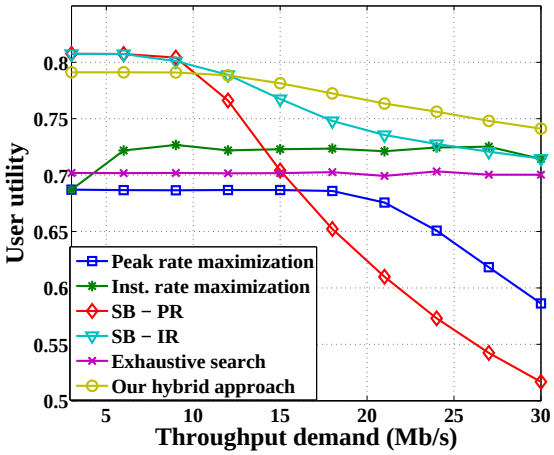


Fig. 28 User utility: Streaming sessions scenario

case of the SB - PR method. However, when the mobile WiMAX gets loaded, users may start to join the LTE technology according to their radio conditions and preferences (*i.e.,* their willingness to pay for better performances). Based on their modulation and coding scheme, as well as on their cost tolerance parameter and decision criteria weights (cf. Table 17), users estimate the utility they can obtain in both available RATs. They then select the technology with the highest expected utility. In fact, users with bad radio conditions that are ready to pay for better performances are the first to start to join the LTE technology. Besides, users with good radio conditions that seek to save up money are the last to join the LTE technology.

Consequently, since users are not proportionally distributed over the two RATs, mobile WiMAX gets overloaded before LTE. Thus, the growth rate of the network utility decreases as the total throughput demand increases (cf. Fig. 27). This means that the average user-perceived throughput decreases. Yet, it remains greater than R_{av}. When some users start joining LTE and so pay more while others, connected to WiMAX, start perceiving lower throughputs, the average user satisfaction also decreases as the total throughput demand increases (cf. Fig. 28).

Furthermore, our hybrid approach and the instantaneous rate maximization method perfectly meet user QoS needs, even at high traffic load. Their network utility, as depicted in Fig. 27, is very close to that of the exhaustive search method, known to be the optimal one with respect to resource utilization. Yet, as shown in Fig. 28, our hybrid approach provides the highest user utility.

In fact, when the instantaneous rate maximization method is used, mobiles select the RAT that offers them the best throughput. Therefore, load balancing is achieved: Mobile WiMAX and LTE are similarly occupied with respect

to their maximum capacity. As a result, the network utility can likely follow the throughput demand increase. On the other hand, when our hybrid approach is employed, the network modulates the broadcasted QoS parameters as a function of its load conditions. It tries to push future arrivals to less loaded RATs, thus enhancing resource utilization. By integrating their needs and preferences, mobiles can avoid oversized decisions and so improve their perceived satisfaction. Typically, at low traffic load, when both RATs can perfectly meet user QoS needs, mobile WiMAX will be preferred since it charges less. This explains why, when using our hybrid method, user utility is constantly higher than when adopting the instantaneous rate maximization method. The latter ignores user preferences (*i.e.,* its willingness to pay for better performances or to save up money) and mainly deals with load balancing. However, because the proportion of users that are connected to the LTE technology is almost constant and the user-perceived throughput is always close to R_{max}, user utility hardly changes as a function of the total throughput demand. On the other side, when using our hybrid method, since the proportion of users that are connected to LTE increases with the total throughput demand, the average user utility decreases since LTE charges more than WiMAX. Yet, it always remains greater than that of the instantaneous rate maximization method.

Moreover, when using the exhaustive search method, the network involves all users at each decision epoch: it considers all possible combinations and selects the one that maximizes its own utility. Since user needs and preferences are ignored, and RATs are not statistically similarly occupied, this network-centric method provides the lowest user utility amongst the instantaneous rate maximization method and our hybrid approach. As a matter of fact, the network seeks to optimize its own utility, regardless of user preferences. In other words, when different combinations lead to the same network utility, they are assumed equivalent. The one that better distributes mobiles over the two RATs has no priority, since it does not improve the network utility defined as the total offered throughput. As a result, the proportion of users that are connected to the LTE technology is statistically higher than those of the instantaneous rate maximization and our hybrid methods, leading to lower user-perceived satisfaction.

To conclude, so as to illustrate the gain from masking network load conditions and only signaling cost and some QoS parameters, we compare our hybrid approach with the SB - IR one. Actually, when using our hybrid method, we can push users to LTE long before WiMAX really gets overloaded. By reducing the broadcasted QoS parameters in WiMAX, even with $S_1 = 0.6$ and $S_2 = 0.95$, future arrivals are encouraged to join LTE much earlier than the SB - IR scenario. Thereby, sessions are better distributed over the two RATs, leading to higher network utility as shown in Fig. 27.

At low traffic load, both methods perfectly meet user QoS needs. Yet, since the proportion of users that are connected to the most expensive RAT (*i.e.,* LTE) is higher when our hybrid approach is used, user-perceived satisfaction is lower than that of the SB - IR method. However, at high throughput demand, because future arrivals start to join LTE much earlier than the SB -

IR case, WiMAX is on average less loaded when using our hybrid approach. As a consequence, WiMAX can better serve its on-going sessions leading to higher user-perceived throughput. Therefore, although mobiles may pay more (*i.e.,* the proportion of users that are connected to LTE is higher), they experience significantly better performances leading to higher satisfaction (Fig. 28). After all, by dynamically tuning QoS parameters, the network enhances resource utilization while mobiles maximize their satisfaction (cf. Fig. 28).

7.2.2 Elastic sessions

We respectively depict in Fig. 29 and 30 the network utility and the average user utility as a function of the total number of users denoted by N_{total}.

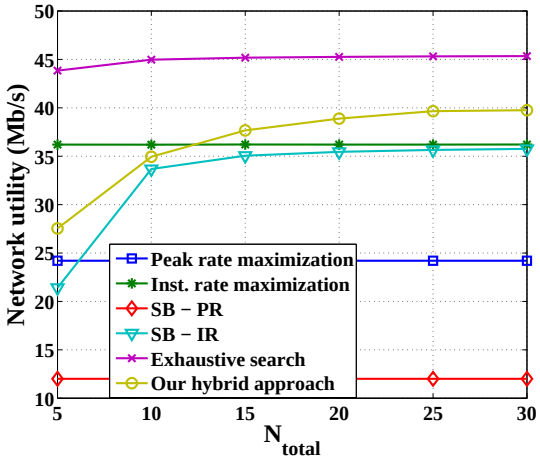


Fig. 29 Network utility: Elastic sessions scenario

When connected alone to a RAT, an elastic session can occupy all of the available resources. However, when several sessions are present, they all share these resources. As a result, the network utility, defined as the total offered throughput, do not usually change as a function of the total number of users N_{total} (cf. Fig 29). Yet, the average user-perceived throughput is reduced.

As in the case of streaming sessions, when the SB - PR method is used, all users are connected to the mobile WiMAX technology regardless of the network load conditions. As shown in Fig. 29, the total offered throughput (*i.e.,* the network utility) is close to 12 Mb/s independently of N_{total}: it actually corresponds to the weighted average total throughput taking into account users with both good and bad radio conditions. However, the average user-perceived throughput linearly decreases with N_{total}, leading to a significant decrease of the user-perceived satisfaction (cf. Fig. 30).

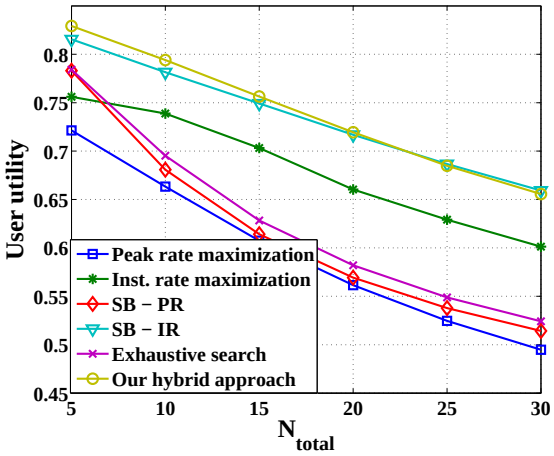


Fig. 30 User utility: Elastic sessions scenario

Moreover, when the peak rate maximization method is adopted, all users select the LTE technology. The network utility is then, on average, higher than that of the SB - PR method. As a consequence, user-perceived throughput is also higher. But, since all users are connected to the most expensive RAT (*i.e.,* LTE), the satisfaction improvement with respect to the perceived throughput criterion fails to offset the satisfaction decrease with respect to the cost criterion. This leads to a lower user-perceived satisfaction in comparison with the SB - PR case (cf. Fig. 30).

Furthermore, when the exhaustive search method is employed, optimal resource utilization is achieved as shown in Fig. 29. Yet, the average user utility is not that interesting. First, when assigning mobiles to the available RATs, this network-centric method does not consider user preferences. It actually ignores user willingness to pay for better performances or to save up money, and only seeks to maximize the network offered throughput. Second, in order to better exploit the available resources, only few users with good radio conditions may be assigned to LTE. The majority, with bad and also good radio conditions, will be connected to mobile WiMAX, all competing for the same resources. As a result, few users connected to LTE will have excellent throughputs that far outweigh their R_c. The others will experience relatively low throughputs that may be well below their R_c. This association optimizes the total offered throughput, but not the user-perceived satisfaction (cf. Fig. 30).

In comparison with the exhaustive search method, mobiles are better distributed over the two RATs when the instantaneous rate maximization method is adopted. In fact, users select the RAT that offers them the best throughput, leading to load balancing as in the streaming case. As a result, mobiles with equivalent radio conditions will have close throughputs regardless of their RAT. Since even users with bad radio conditions may be connected to LTE,

the network utility is on average lower than that of the exhaustive search method known to be the optimal one. However, because on average perceived throughputs better meet user needs (*i.e.,* their R_c), the user utility is significantly higher than that of the exhaustive search approach.

On the other hand, when the SB - IR method is used, mobile users combine their needs and preferences with the network load conditions so as to select their best RAT. At low traffic load (typically for $N_{total} = 5$), more users select the mobile WiMAX technology in comparison with the instantaneous rate maximization method. When WiMAX can meet user needs very well, it charges them less. Occasionally, based on the current load conditions, a user with bad radio conditions, that is ready to pay for better performances, would select the LTE technology. As N_{total} increases, more users including those with good radio conditions start to join LTE, leading to higher network utility. The latter remain almost constant at medium and high load conditions. On average, it is slightly lower than that of the instantaneous rate maximization method. Yet, since selection decisions take into account user needs and preferences, typically their cost considerations, the user utility is significantly better than that of the instantaneous rate maximization method.

Lastly, by masking network load conditions and only signaling some cost and QoS parameters, our hybrid approach drives user decisions in a way to enhance resource utilization. At low traffic load, more users typically those with bad radio conditions, that are ready to pay, select LTE. This leads to a higher network utility in comparison with the SB - IR method where, as explained before, users may occasionally join LTE (cf. Fig. 29). As a result, and although users pay on average more, they experience higher satisfaction since they have quite better throughput.

As N_{total} increases, QoS parameters are reduced with $S_1 = 0.3$ and $S_2 = 0.8$. As a consequence, future arrivals are encouraged to join LTE much earlier than the SB - IR case. However, users with good radio conditions that seek to save up money are the last to start joining LTE. In comparison with the SB - IR method, most users that are connected to WiMAX have good radio conditions, and more users with both good and bad radio conditions are connected to LTE. This leads to higher total offered throughput, as shown in Fig. 29. Yet, the user utility is pretty close to that of the SB - IR scenario, since users having better performances pay on average more.

To wrap up, in comparison with different RAT selection schemes, including network-centric, hybrid and mobile-terminal-centric approaches, simulation results prove the efficiency of our hybrid approach in enhancing resource utilization and maximizing user satisfaction. In the streaming sessions scenario, it optimizes the total offered throughput and maximizes the average user utility, except at low traffic load where the non-realistic SB - IR method provides higher user satisfaction. Also, in the elastic sessions scenario, our hybrid approach significantly enhances resource utilization and maximizes user utilities in comparison with various hybrid and mobile-terminal-centric methods. Furthermore, compared with the exhaustive search method, known to

be the optimal one with respect to resource utilization, our hybrid approach provides significantly higher user satisfaction.

8 Conclusion

In this paper, we addressed the radio access technology selection, a key common radio resource management functionality in heterogeneous wireless networks. We identified the need and proposed a hybrid approach that combines benefits from both network-centric and mobile-terminal-centric methods. As a matter of fact, the network information that is periodically broadcasted assists mobile users in their decisions: mobiles select their RAT based on their needs and preferences as well as on the cost and partial QoS parameters signaled by the network. On the one hand, by broadcasting appropriate decisional information, the network tries to globally control user decisions in a way to meet operator objectives (*e.g.,* enhance resource utilization). On the other hand, mobile users make their decisions so as to maximize their own utility. Selection decisions then integrate operator objectives and user needs and preferences, without unduly complicating the network.

We also presented a satisfaction-based multi-criteria decision-making method, that mobiles use to evaluate the different alternatives and then select their RAT. In comparison with existing methods, our algorithm meets user needs (*e.g.,* traffic class, throughput demand, cost tolerance), avoiding oversized and undersized decisions. Furthermore, we introduced two heuristic methods, namely the staircase and the slope tuning policies, to dynamically derive network information. While QoS parameters are modulated as a function of the load conditions, radio resources are efficiently exploited.

When users do not cooperate neither with each other nor with the network, they have no information regarding the global network state. As a result, their selection decisions may be in no one long-term interest, leading to performance inefficiency. Moreover, when the network takes selection decisions transparently to end-users, resource utilization is optimized. Yet, individual user needs and preferences are not efficiently met, leading to relatively low user satisfaction. However, when our hybrid approach is used, the network partially cooperates with mobiles assisting them in their decisions. The network actually masks its load conditions and only signals cost and some QoS parameters. This decisional information guides user decisions in a way to enhance resource utilization. Besides, since user needs and preferences are also integrated, selection decisions maximize user satisfaction.

We proved as well the efficiency of masking network load conditions, and only signaling cost and some QoS parameters, in enhancing resource utilization and user satisfaction. In fact, our hybrid approach outperforms non-realistic methods, where mobiles have a perfect knowledge of the network state (*i.e.,* numbers of users connected to available RATs). So, to conclude, when operator objectives are implicitly integrated within signaled QoS parameters, radio resources are better utilized and user satisfaction is maximized.

Finally, compared with various hybrid and mobile-terminal-centric methods, our hybrid approach maximizes the total offered throughput and the average user satisfaction. Also, compared with the optimal exhaustive search method, our approach provides significantly higher user utility.

References

1. IEEE Standard for Architectural Building Blocks Enabling Network-Device Distributed Decision Making for Optimized Radio Resource Usage in Heterogeneous Wireless Access Networks. IEEE Std 1900.4-2009 (2009)
2. 3GPP TS 32.521: Telecommunication Management; Self-Organizing Networks (SON) Policy Network Resource Model (NRM) Integration Reference Point (IRP); Requirements (2010)
3. Aryafar, E., Keshavarz-Haddad, A., Wang, M., Chiang, M.: RAT Selection Games in HetNets. In: Proc. IEEE Conference on Computer Communications (INFOCOM) (2013)
4. Bari, F., Leung, V.C.: Automated Network Selection in a Heterogeneous Wireless Network Environment. IEEE Networks **21**(1), 34 – 40 (2007)
5. Chamodrakas, I., Martakos, D.: A Utility-Based Fuzzy TOPSIS Method for Energy Efficient Network Selection in Heterogeneous Wireless Networks. Applied Soft Computing **12**(7), 1929 – 1938 (2012)
6. Coupechoux, M., Kelif, J.M., Godlewski, P.: Network Controlled Joint Radio Resource Management for Heterogeneous Networks. In: Proc. IEEE Vehicular Technology Conference (VTC Spring) (2008)
7. Coupechoux, M., Kelif, J.M., Godlewski, P.: SMDP Approach for JRRM Analysis in Heterogeneous Networks. In: Proc. European Wireless Conference (EW) (2008)
8. Dhahri, C., Ohtsuki, T.: Learning-Based Cell Selection Method for Femtocell Networks. In: Proc. IEEE Vehicular Technology Conference (VTC Spring) (2012)
9. Dhahri, C., Ohtsuki, T.: Q-learning Cell Selection for Femtocell Networks: Single- and Multi-user Case. In: Proc. IEEE Global Communications Conference (GLOBECOM) (2012)
10. Edell, R., Varaiya, P.: Providing Internet Access: What We Learn from INDEX. IEEE Network **13**(5), 18 – 25 (1999)
11. El Helou, M., Ibrahim, M., Lahoud, S., Khawam, K.: Optimizing Network Information for Radio Access Technology Selection. In: Proc. IEEE Symposium on Computers and Communications (ISCC) (2014)
12. El Helou, M., Ibrahim, M., Lahoud, S., Khawam, K., Mezher, D., Cousin, B.: A Network-Assisted Approach for RAT Selection in Heterogeneous Cellular Networks. IEEE Journal on Selected Areas in Communications **33**(6), 1055 – 1067 (2015)
13. Falowo, O., Chan, H.: RAT Selection for Multiple Calls in Heterogeneous Wireless Networks Using Modified TOPSIS Group Decision-Making Technique. In: Proc. IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC) (2011)
14. Gazis, V., Alonistioti, N., Merakos, L.: Toward a Generic "Always Best Connected" Capability in Integrated WLAN/UMTS Cellular Mobile Networks (and Beyond). IEEE Wireless Communications **12**(3), 20 – 29 (2005)
15. Giupponi, L., Agustí, R., Pérez-Romero, J., Sallent, O.: A Framework for JRRM with Resource Reservation and Multiservice Provisioning in Heterogeneous Networks. Mobile Networks and Applications **11**(6), 825 – 846 (2006)
16. Gozalvez, J., Lucas-Estañ, M.C., Sanchez-Soriano, J.: Joint Radio Resource Management for Heterogeneous Wireless Systems. Wireless Networks **18**(4), 443 – 455 (2012)
17. Gueguen, C., Baey, S.: A Fair Opportunistic Access Scheme for Multiuser OFDM Wireless Networks. EURASIP Journal on Wireless Communications and Networking (2009) (2009)
18. Gustafsson, E., Jonsson, A.: Always Best Connected. IEEE Wireless Communications **10**(1), 49 – 55 (2003)

19. Hardin, G.: The Tragedy of the Commons. Science Journal (1968)
20. Ibrahim, M., Khawam, K., Tohme, S.: Network-Centric Joint Radio Resource Policy in Heterogeneous WiMAX-UMTS Networks for Streaming and Elastic traffic. In: Proc. IEEE Wireless Communications and Networking Conference (WCNC) (2009)
21. Ibrahim, M., Khawam, K., Tohme, S.: Congestion Games for Distributed Radio Access Selection in Broadband Networks. In: Proc. IEEE Global Communications Conference (GLOBECOM) (2010)
22. Khawam, K.: The Modified Proportional Fair Scheduler. In: Proc. IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC) (2006)
23. Khawam, K., Ibrahim, M., Cohen, J., Lahoud, S., Tohme, S.: Individual vs. Global Radio Resource Management in a Hybrid Broadband Network. In: Proc. IEEE International Conference on Communications (ICC) (2011)
24. Khawam, K., Marinca, D.: Size-based Proportional Fair Scheduling. In: Proc. IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC) (2010)
25. Kumar, D., Altman, E., Kelif, J.M.: User-Network Association in a WLAN-UMTS Hybrid Cell: Global & Individual Optimality. Rapport de recherche RR-5961, INRIA (2006)
26. Lucas-Estañ, M.C., Gozalvez, J., Sanchez-Soriano, J.: Integer Linear Programming Optimization of Joint RRM Policies for Heterogeneous Wireless Systems. Computer Networks **56**(1), 112 – 126 (2012)
27. Moety, F., Ibrahim, M., Lahoud, S., Khawam, K.: Distributed Heuristic Algorithms for RAT Selection in Wireless Heterogeneous Networks. In: Proc. IEEE Wireless Communications and Networking Conference (WCNC) (2012)
28. Naja, R., El Helou, M., Tohmé, S.: WiMAX Double Movable Boundary Scheme in the Vehicle to Infrastructure Communication Scenario. Wireless Personal Communications **67**(2), 387 – 413 (2012)
29. Nguyen-Vuong, Q.T., Agoulmine, N., Cherkaoui, E., Toni, L.: Multicriteria Optimization of Access Selection to Improve the Quality of Experience in Heterogeneous Wireless Access Networks. IEEE Transactions on Vehicular Technology **62**(4), 1785 – 1800 (2013)
30. Niyato, D., Hossain, E.: Dynamics of Network Selection in Heterogeneous Wireless Networks: An Evolutionary Game Approach. IEEE Transactions on Vehicular Technology **58**(4), 2008 – 2017 (2009)
31. Pérez-Romero, J., Gelabert, X., Sallent, O.: Radio Resource Management for Heterogeneous Wireless Access Networks. In: E. Hossain (ed.) Heterogeneous Wireless Access Networks: Architectures and Protocols, chap. 5, pp. 133 – 165. Springer (2008)
32. Premkumar, K., Kumar, A.: Optimum Association of Mobile Wireless Devices with a WLAN-3G Access Network. In: Proc. IEEE International Conference on Communications (ICC) (2006)
33. Sang, A., Wang, X., Madihian, M., Gitlin, R.D.: Coordinated Load Balancing, Handoff/Cell-Site Selection, and Scheduling in Multi-Cell Packet Data Systems. Wireless Networks **14**(1), 103 – 120 (2008)
34. Singh, J.P., Alpcan, T., Agrawal, P., Sharma, V.: A Markov Decision Process based Flow Assignment Framework for Heterogeneous Network Access. Wireless Network **16**(2), 481 – 495 (2010)
35. Stevens-Navarro, E., Wong, V.: Comparison Between Vertical Handoff Decision Algorithms for Heterogeneous Wireless Networks. In: IEEE Vehicular Technology Conference (VTC Spring) (2006)
36. Tabrizi, H., Farhadi, G., Cioffi, J.: Dynamic Handoff Decision in Heterogeneous Wireless Systems: Q-learning Approach. In: Proc. IEEE International Conference on Communications (ICC) (2012)
37. Toskala, A., Holma, H., Kolding, T., Mogensen, P., Pedersen, K., Reunanen, J.: High-Speed Downlink Packet Access. In: H. Holma, A. Toskala (eds.) WCDMA FOR UMTS - HSPA Evolution and LTE, Fifth Edition, chap. 12, pp. 353 – 389. Wiley (2010)
38. Wang, L., Binet, D.: Mobility-Based Network Selection Scheme in Heterogeneous Wireless Networks. In: Proc. IEEE Vehicular Technology Conference (VTC Spring) (2009)
39. Yu, F., Krishnamurthy, V.: Efficient Radio Resource Management in Integrated WLAN/CDMA Mobile Networks. Telecommunication Systems **30**(1-3), 177 – 192 (2005)

40. Zhang, F., Yan, Y., Ahmad, A.: Pricing for Efficient Usage in Wired and Wireless Networks. In: Proc. International Telecommunications Network Strategy and Planning Symposium (NETWORKS) (2004)
41. Zhang, W.: Handover Decision Using Fuzzy MADM in Heterogeneous Networks. In: Proc. IEEE Wireless Communications and Networking Conference (WCNC) (2004)
42. Zhang, X., Jin, H., Ji, X., Li, Y., Peng, M.: A separate-SMDP Approximation Technique for RRM in Heterogeneous Wireless Networks. In: Proc. IEEE Wireless Communications and Networking Conference (WCNC) (2012)
43. Zhu, L., Yu, F., Ning, B., Tang, T.: Cross-Layer Handoff Design in MIMO-Enabled WLANs for Communication-Based Train Control (CBTC) Systems. IEEE Journal on Selected Areas in Communications **30**(4), 719 – 728 (2012)
44. Zhu, L., Yu, F.R., Ning, B., Tang, T.: Handoff Management in Communication-Based Train Control Networks Using Stream Control Transmission Protocol and IEEE 802.11p WLANs. EURASIP Journal on Wireless Communications and Networking **2012**(1), 211 – 226 (2012)